Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 07 Septembre 2022

Par : Mulah MORIAH
Titre : Mesure et mitigation des biais : vers une tarification non-vie réellement équitable
Sous-titre : Assurer l'équité, un enjeu sociétal et stratégique.

Confidentialité : Non

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membre présent du jury de l'Institut des Actuaires :
Alexandre YOU
Fabrice HAMON
Signature :

Entreprise :
OPTIMIND
Signature :

Membres présents du jury de l'EURIA :
Pierre AILLIOT

Directeur de mémoire en entreprise :
André GRONDIN
Signature :

Invité :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

L'assurance non-vie est un marché très concurrentiel et réglementé dans lequel la tarification se trouve à la croisée de plusieurs chemins. En effet, l'assureur doit utiliser tous les outils statistiques et les données à sa disposition pour construire les meilleurs tarifs possibles. Dans le même temps, ses primes doivent être alignées avec la stratégie de l'entreprise et prendre en compte la concurrence.

Du fait du rôle important qu'occupe l'assurance dans la société, les primes sont aussi scrutées par les régulateurs. Elles doivent être transparentes, explicables et éthiques. Ainsi, la tarification n'est pas que statistique, elle porte aussi des enjeux stratégiques et sociétaux. Ces différents enjeux peuvent pousser l'assureur à proposer des primes plus équitables par rapport à une variable donnée. Par exemple, la règlementation oblige les assureurs à présenter des primes équitables suivant le genre des assurés ou des groupes mutualistes, étant donné leurs stratégies, proposent des primes équitables suivant l'âge. Dans d'autres domaines de l'assurance, des variables comme la présence de maladies graves ou de handicap commencent à être soumises à la notion d'équité.

Peu importe la raison pour laquelle un acteur de l'assurance doit présenter une tarification plus équitable par rapport à une variable, il doit être capable de définir, mesurer puis mitiger le biais éthique de sa tarification tout en préservant sa cohérence et sa performance. Ces travaux ont pour objectif de fournir ces différents outils et de les tester sur un cas d'usage réaliste d'assurance automobile.

Mots clefs: Équité, Mesures, Mitigation, Apprentissage statistique, Transparence, Assurance IARD

Abstract

Non-life insurance is a highly competitive and regulated market in which pricing is at a crossroads. Indeed, the insurer must use all the statistical tools and data at its disposal to build the best possible rates. At the same time, its premiums must be aligned with the company's strategy and take into account competition.

Because of the important role played by insurance in society, premiums are also scrutinized by regulators. They must be transparent, explainable and ethical. Thus, pricing is not only statistical, it also carries strategic and societal issues. These different issues can push the insurer to offer fairer premiums in relation to a given variable. For example, the regulations require insurers to present fair premiums according to the gender of the insured or mutualist groups, given their strategies, offer fair premiums according to age. In other areas of insurance, variables such as the presence of serious illness or disability are beginning to be subject to the notion of fairness.

Regardless of the reason why an insurance player must present fairer pricing in relation to a variable, it must be able to define, measure and then mitigate the ethical bias of its pricing while preserving its consistency and performance. The objective of this work is to provide these different tools and to test them on a realistic car insurance use case.

Keywords: Fairness, Measurement, Mitigation, Statistical learning, Transparency, P&C Insurance

Remerciements

J'aimerais commencer par remercier Optimind de m'avoir accueilli pour ce stage de fin d'études, remerciant tout particulièrement André Grondin et Bénédicte Agbré, mes tuteurs pour leur encadrement et leur apport. Je remercie aussi Guillaume Besson et Geoffroy Lambolez pour leur aide précieuse.

Je remercie Franck Vermet, mon tuteur académique ainsi que toute l'équipe enseignante et administrative de l'EURIA pour leur encadrement, leur soutien et leur engagement tout au long de la formation et de la réalisation du mémoire.

Je remercie aussi mes précieux amis pour leur aide et leur soutien. Remerciements chaleureux à Lauriane et Camille sans qui la rédaction de ce mémoire aurait été beaucoup plus difficile.

J'aimerais terminer en remerciant mes parents qui ont consenti d'énormes sacrifices pour faire de moi ce que je suis, pour leur soutien et amour sans limites. Je leur dédie l'accomplissement de ces années d'études supérieures.

All things are possible to him
that believeth.

Note de synthèse

Contexte

L'assurance est caractérisée par l'inversion de son cycle de production : l'assureur demande une prime fixée à la souscription en échange d'une couverture contre des risques dont la réalisation et le montant sont aléatoires. Cette inversion met un accent sur le caractère statistique qui entoure la tarification. Celle-ci doit être cohérente avec la théorie statistique pour permettre l'estimation et la couverture des phénomènes aléatoires.
Au-delà de ces considérations statistiques, les primes représentent le juste prix du service d'assurance, et comme tout prix, il encapsule de nombreux enjeux stratégiques. Ainsi, un acteur du marché de l'assurance doit proposer des prix compétitifs, alignés avec la stratégie et la communication de l'entreprise mais aussi utiliser des outils adaptés aux différents processus de distribution.

Ces dernières années ont été marquées par l'utilisation des algorithmes dit de machine learning et des réseaux de neurones ainsi que des données massives. Pour cause, les avancées scientifiques mais surtout l'augmentation de la puissance de calcul, son accessibilité et la grande quantité de données en circulation. Ces données et algorithmes servent d'outils d'aide à la prise de décision, ils permettent de segmenter les assurés, de comprendre le risque et tous les différents facteurs qui l'entoure. Les acteurs doivent donc intégrer ces nouveaux éléments pour ne pas perdre en compétitivité.
Toutefois, l'utilisation de ces données massives et d'algorithmes complexes a rapidement mis sur la table le sujet de la transparence ; les primes et les décisions doivent être explicables et équitables. En effet, les enjeux sont grands, il ne s'agit pas d'algorithmes permettant de choisir un film mais d'algorithme fixant le coût de l'accès aux services d'assurance pour les segments de la société.
Ces sujets d'équité et d'éthique font partie entière de nos sociétés et des réflexions depuis des siècles. Bien que sujette à interprétation, l'équité peut se définir comme étant la capacité à mettre des individus sur un pied d'égalité en tenant compte des différences qui existent entre-eux. C'est par principe d'équité que des règles telles que la gender directive ont vu le jour. Les réformes sur l'accès à l'assurance emprunteur sont aussi une forme d'application de l'équité. Ainsi, les acteurs du secteur de l'assurance peuvent devoir construire des primes plus équitables par rapport à des variables dites sensibles ou protégées. Ces contraintes d'équité peuvent émaner de la règlementation comme c'est

le cas avec le genre ou des objectifs commerciaux et stratégiques comme c'est le cas avec l'âge dans certaines entreprises.

Dans ces travaux, la notion de biais traitée est le biais dans le sens de discrimination, elle désigne l'effet indésirable d'une variable sensible sur une variable d'intérêt, par exemple l'effet du genre sur la prime dans le cadre imposé par la gender directive.
Depuis 2016, de nombreux travaux de recherches pointent du doigt les discriminations présentes dans des outils d'aide à la décision : l'outil de mesure du risque de récidive aux États-unis, les algorithmes de Google images, de traitement de candidatures Amazon, les algorithmes d'octroi de prêts et de financements et bien d'autres encore.
Depuis longtemps, la solution utilisée est de contourner le problème sans le résoudre du point de vue éthique. Une situation simplifiée dans laquelle une variable V corrélée avec le genre est utilisée pour déterminer la prime a été mise en scène. Le genre n'est pas pris en compte dans les estimations, il s'agit de l'équité par omission. Toutefois, la présence de corrélation entre le genre et la variable explicative non sensible conduit à un effet indirect du genre sur les primes estimées. Ainsi, bien qu'en apparence le genre soit retraité ou supprimé du modèle, sa relation avec les autres variables permet de conserver son effet sur la variable cible. Cela est dû au fait que les retraitements ne permettent de traiter que l'effet direct des variables sensibles et non leurs effets indirects alors que ces variables exercent une influence non négligeable sur les distributions des autres variables explicatives. L'âge, tout comme le genre ou le handicap affecte les choix d'activités, la prise de risque, les choix de produits, etc.
Il est donc indispensable de pouvoir dans un premier temps définir et mesurer l'équité des modèles construits.

Définition et mesure de l'équité

La littérature sur l'étude du biais dans les modèles est assez récente. Les références et les consensus ne sont pas explicites comme le décrivent des chercheurs américains en avril 2021 : "La croissance rapide de ce nouveau domaine a conduit à des motivations, terminologies et notations extrêmement incohérentes, présentant un sérieux défi pour le catalogage et la comparaison des définitions". Dans la 12eme édition du Scientific Report de mars 2022 sur la science des données et l'intelligence artificielle, les auteurs parlent du zoo des définitions de l'équité : "le chercheur ou le praticien abordant cette facette du machine learning pour la première fois peut facilement se sentir confus et en quelque sorte perdu dans ce zoo de définitions. Ces multiples définitions saisissent différents aspects du concept d'équité mais, au summum de nos connaissances, il n'y a toujours pas de compréhension claire du paysage global où vivent ces mesures."

Ces travaux proposent une revue de la littérature sur le sujet, une littérature quasi exclusivement anglophone. La revue ne sera pas exhaustive mais assez claire et complète pour être utilisée pour résoudre différents problèmes d'équité en assurance. De plus, le cas de la classification est prépondérant dans la littérature, ces travaux proposent donc

des méthodes et des adaptations pour permettre la prise en compte de l'équité dans le cas continue qu'impose la tarification.

L'équité se comprend intuitivement comme étant l'absence de dépendance entre la variable sensible et les variables d'intérêts. En effet, l'indépendance implique qu'aucune relation directe ou indirecte n'existe entre ces variables et le biais ne peut être présent. Cette définition basée sur l'indépendance n'est en réalité qu'un des aspects de l'équité. Il faut d'abord distinguer le cadre des observations du cadre intégrant la causalité. Le cadre causal bien qu'intéressant n'est pas traité en détails car les outils statistiques pour faire de la causalité sont encore trop restreints. Ensuite, il faut discerner équité de groupe et équité individuelle. L'équité de groupe ou équité statistique stipule que certaines statistiques doivent être les mêmes entre les individus des différents groupes formés par la variable sensible, il s'agit de l'approche classique. L'équité individuelle quant à elle spécifie que des individus semblables doivent être traités de la même manière. Cette approche est la plus intuitive mais aussi la plus difficile et la plus coûteuse à mettre en place du fait qu'elle soit liée à la causalité.

Enfin, plusieurs métriques ont été définies pour tenter de quantifier le niveau d'équité. Six mesures sont retenues pour les applications, des mesures classiques, telles que le ratio des moyennes et le tau de Kendall et des mesures de différences entre distributions de probabilités telles que le test de Kolmogorov-Smirnov et la divergence de JS. En plus de ces mesures, deux mesures occupent une place centrale dans les travaux, ce sont le HGR KDE et l'adaptation du flip-test.

HGR KDE : cette métrique est la plus solide théoriquement. Elle permet de mesurer tout type de dépendances entre tout type de variables. Deux versions ont été implémentées ; la version basée sur les réseaux neuronaux et la version utilisant les noyaux (KDE). La version KDE a été retenue car elle converge deux fois plus rapidement que celle basée sur les réseaux neuronaux. De plus, le choix de la structure optimale du réseau de neurones est délicat. Une structure plus légère permettrait une convergence plus rapide mais éventuellement des résultats moins fiables.

Adaptation flip-test : cette métrique permet de définir l'équité individuelle sans nécessiter les constructions d'un graphe et d'un modèle causal. Cela est possible en contrepartie de l'hypothèse qu'une distance ou qu'un algorithme permette de mesurer la proximité entre les individus de la base de données. Initialement construite pour une variable cible binaire, elle a été adaptée dans ces travaux en utilisant un algorithme de k plus proches voisins comme mesure de proximité. Ainsi, il est possible de comparer la prédiction d'un individu à celle des individus du genre opposé les plus proches. La qualité de la notion de proximité fournie par le modèle de k plus proches voisins construit est la principale limite de cette approche. Pour contrôler cette limite, un hyperparamétrage et une sélection de variables sont menés. La distance choisie et son paramétrage ainsi que le nombre de voisins sont optimisés. Le nombre de variables et les variables retenues pour construire le modèle sont aussi optimisés, l'objectif étant d'obtenir le modèle de k

plus proches voisins conduisant à la plus petite distance entre les individus et au plus faible écart de primes. Ainsi, sur ces différents paramètres, une recherche par grille est effectuée et les résultats sont vérifiés sur une base de test. Cette approche fournit une solution satisfaisante avec un coût opérationnel significativement plus faible que celui qu'engendrerait une étude causale.

Après la détection du biais, il est indispensable de tenter de le mitiger tout en préservant les performances et la cohérence des modèles construits.

Mitigation du biais

La mitigation du biais est le fait de mener des actions sur les modèles, les données ou les résultats dans le but de réduire le biais entre la variable sensible et l'estimation de la variable d'intérêt. Comme dans le cas des mesures, aucun consensus n'existe sur les approches de mitigation du biais. Le champ de recherches est en pleine expansion et les approches proposées dans la littérature sont dans certains cas trop rattachées au cas d'usage qu'elles traitent. Le cas Y binaire est ici encore le cas de prédilection. Dans ces travaux, des méthodes utilisables pour le cas de la régression sont proposées. La mitigation peut être pratiquée avant, pendant ou après la modélisation.

Ante modélisation : ces méthodes reviennent à transformer les données dans le but de réduire le biais tout en préservant les informations pertinentes. Les méthodes implémentées sont la suppression totale des variables corrélées à la variable sensible, la suppression des corrélations linéaires et une adaptation de la méthode fair-SMOTE. Initialement construite pour le cas binaire, cette méthode est adaptée pour répondre au besoin de la mitigation en tarification non-vie.

∗ *Avantages :* ces méthodes ont l'avantage de permettre de garder intact tout le processus de modélisation. Elles sont en général simples à implémenter et moins coûteuses en temps de calculs.
∗ *Inconvénients :* elles peuvent conduire à des pertes d'informations significatives. Les variables peuvent être étroitement liées entre elles et ne pas permettre à la fois la suppression des biais et la conservation des informations. De plus, il faudra s'assurer que le reste du processus de modélisation n'introduise pas de biais.

Pendant modélisation : ces méthodes consistent à intégrer des contraintes d'équité directement dans la phase de calibration des modèles. Deux méthodes sont implémentées, l'exponentiated gradient qui tire son nom de la méthode de la théorie des jeux sur laquelle elle s'appuie, et sa version de recherche par grille.

∗ *Avantages :* ces méthodes utilisent les informations sur la variable sensible dans le but de trouver le meilleur équilibre entre performance et équité. Théoriquement, elles sont plus à même de conduire au meilleur arbitrage possible.

∗ *Inconvénients :* elles sont difficiles à mettre en place et à généraliser. De plus, même après une implémentation réussie, leur convergence n'est pas assurée et les temps de calculs peuvent être exponentiels.

Post modélisation : elle consiste à transformer les prédictions obtenues en sortie des modèles dans le but de les rendre plus équitables. Le premier exemple auquel il est facile de penser est celui de la régression logistique. Les probabilités obtenues permettent de modifier les seuils de décision avec pour conséquence d'impacter le comportement du modèle vis-à-vis des différentes classes. Les méthodes post modélisation présentes dans la littérature sont donc des méthodes de modifications de frontières de décision prenant en compte des mesures d'équité. Ces approches ne sont pas applicables au cas continu car il n'y a pas de frontières de décision. De plus, aucune publication appliquant et testant la mitigation post modélisation dans ce cas n'a pu être trouvée.

Des études ont donc été menées dans le but de proposer une approche de mitigation dans le cas continu. Pour pouvoir imposer l'équité après le calcul des primes, il faudrait pouvoir définir ce qu'est une prime avantageuse et une prime désavantageuse. Toutefois, cela n'est pas possible dans l'absolu car une prime avantageuse dépend des caractéristiques de l'assuré et du risque porté. C'est à la suite de ce constat que l'approche individuelle « redistribution équitable » a été proposée. Elle exploite l'adaptation du flip-test comme définition du biais pour segmenter la prime en une partie équitable et un biais. Ce biais est ensuite redistribué aux individus de la base de manière itérative, dans le but de converger vers un état dans lequel, au vu des nouvelles primes, le biais mesuré soit le plus faible possible.

∗ *Avantages :* ces méthodes ne nécessitent pas de recalibrer les modèles, elles ont donc des temps de calculs plus courts. Les sorties obtenues ne peuvent plus être contaminées par du biais car ce sont les résultats finaux.
∗ *Inconvénients :* les résultats sont dépendants de la qualité des modèles construits, la mitigation sur des mauvaises prédictions n'est pas très cohérentes.

Application à la tarification automobile

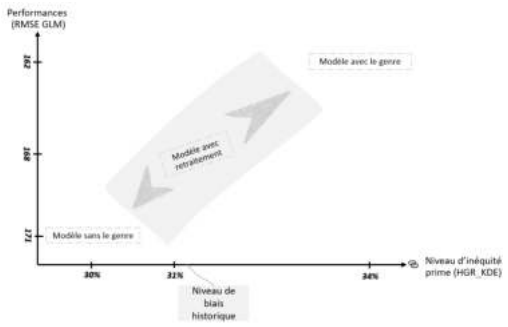
Pour cette application, une garantie bris de glace est tarifée en prenant en compte tout au long du processus les contraintes statistiques mais aussi métiers et opérationnelles. Les données ont été traitées et analysées, les variables retraitées et discrétisées dans le but de construire une base de modélisation. Une base décrivant les véhicules a aussi été intégrée et un zonier a été construit. La RMSE et le ratio de charges totales sur primes totales, noté S/P, ont entre autres été utilisés pour évaluer les modèles construits à l'aide du GLM, du random forest et du gradient boosting. Ces modèles ont été optimisés puis validés en utilisant des méthodes d'interprétabilité.

Les modèles GLM ont été retenus à la suite de cette phase de tarification. En effet, bien que les modèles boîtes noires puissent obtenir de meilleures performances, les GLM

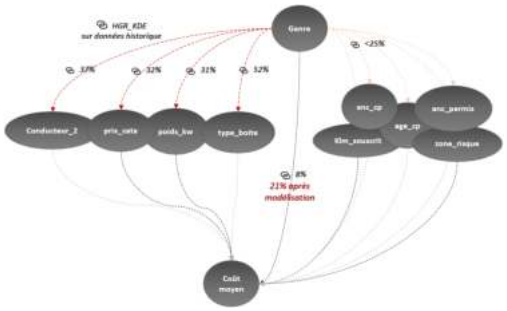
restent les plus répandus car ils sont simples à interpréter et à intégrer dans les outils de tarification en production. Ainsi, la légère hausse de performances obtenue à l'aide du modèle de random forest ne suffit pas pour le sélectionner du fait des coûts opérationnels que cela engendrerait. Toutefois, à chaque étape, les calculs sont effectués sur les trois modèles, et en général il n'y a pas d'écarts significatifs. De plus, le modèle de prime pure est préféré à un modèle multiplicatif coût × fréquence car il est plus performant au sens des métriques définies. Les modèles de références sont construits en omettant le genre. Une fois construits, ces travaux vont plus loin en tentant de mesurer le biais provenant du genre et d'ensuite mitiger ce biais.

Mesure du biais

A l'aide des six mesures de dépendance retenues pour ce cas pratique, les biais sont mesurés sur les données historiques et sur les résultats après modélisation. Pour des raisons de simplicité, les résultats sur les primes pures sont présentés en priorité dans cette note, les tendances observées étant les mêmes pour les coûts moyens et les fréquences. La figure 2a résume les résultats des mesures pour le modèle de prime.



(a) Arbitrage entre performance et équité pour le modèle de prime.



(b) Effets directs et indirects du genre sur le coût moyen.

FIGURE 1 – Quelques résultats sur la mesure des biais.

Les différentes mesures ont permis de détecter du biais dans les données historiques, toutefois, vu qu'il n'existe pas d'antécédents sur le sujet de la mesure, il est difficile de tirer des seuils et de définir le niveau de significativité de ces biais.

Un premier modèle avec le genre est ensuite construit, il conduit a des niveaux de biais plus élevés pour chacune des variables d'intérêts. Le biais dans le modèle de coût est multiplié par 2,6 et passe de 8% à 21%. Ainsi, le biais a non seulement été appris par le modèle mais, il a aussi été amplifié. Les analyses ont permis de se rendre compte que le genre occupe une place importante pour les mesures d'importances des variables. En plus de cela, la variable genre est liée aux autres variables significatives des différents modèles. L'idée est que ces interdépendances avec le genre ont conduit à l'amplification de son effet sur les variables d'intérêts.

Le biais est donc ensuite mesuré sur un second modèle ne contenant pas le genre. Les niveaux de biais atteints sont proches des niveaux obtenus dans l'historique de données. Des mesures de dépendance comme le tau de kendall qui détectait très peu de dépendance avant la modélisation en détecte 2 à 3 fois plus après la modélisation. Ainsi, malgré l'absence du genre, il exerce toujours une influence sur les sorties des modèles. Comme dans l'exemple simplifié, la suppression du genre permet de supprimer son effet direct mais ne traite pas son effet indirect à travers les autres variables explicatives. La figure 2b illustre ces différents effets pour le coût moyen.
Dans des modèles multivariés comme ceux construits ici, ces distinctions entre genres peuvent provenir de combinaisons entre plusieurs variables. Par exemple, une combinaison puissance de véhicule, type de boîte et zone de risque peut être très déséquilibrée en termes de répartition homme-femme. A titre d'illustration, en croisant zonier et prix des véhicules, des déséquilibres significatifs apparaissent entre les genres dans le jeu de données. 87% des individus ayant un véhicule coûtant plus de 30000€ et vivant en zone 12 sont des hommes. Ainsi, toute différence significative de prime entre ce segment et les autres impacte fortement les hommes.

Dans le but de vérifier que ces niveaux de biais ne proviennent pas d'une mauvaise prise en compte du genre dans la tarification initiale, un modèle retraitant le genre à la sortie des modèles a été construit. Il conduit sensiblement aux mêmes résultats, le retraitement ne permettant pas de résoudre le problème provoqué par l'interdépendance. Des tests ont permis de montrer que les variables explicatives pouvaient permettre de prédire le genre en utilisant des modèles de classification. Il apparaît donc que les variables puissent reconstruire le genre et donc conserver son effet sur les prédictions. Dans le but d'obtenir des modèles plus équitables, il est indispensable d'appliquer des méthodes de mitigation plus sensibles à l'équité.

Mitigation du biais

Les différentes approches de mitigations sont testées en tenant compte de l'arbitrage entre performance et équité, suivant l'arbitrage que l'acteur décide de faire, les meilleurs modèles peuvent être modifiés. La mitigation est effectuée directement sur le modèle de prime pure. La notion de modèle non dominé est définie dans le but de permettre la suppression des scénarios non concluants. Un modèle est dit non dominé quand il n'existe aucun autre modèle ayant à la fois une meilleure performance et une meilleure équité que lui. Ainsi, suivant l'arbitrage consenti par l'assureur, un modèle non dominé peut être le meilleur et ne pas l'être suivant un autre arbitrage. Le modèle ne contenant pas le genre est utilisé comme référence tant pour la performance que pour l'équité. Pour chacune des méthodes, des sensibilités au choix des hyperparamètres et des choix de modélisation sont mis en place. Les résultats de ces mitigations sont compilés dans le tableau 4.22. Ils proviennent des cinq méthodes dont les résultats, les apports et les limites sont présentés ci-dessous.

Métriques	Modèle de référence	Suppression totale	Suppression corrélation	Adaptation fair-SMOTE	Exponentiated gradient	Redistribution équitable
HGR KDE	29,71%	19%	26,50%	28,83%	31,22%	30,08%
RMSE	171,04	177,6	169,6	171,61	171,25	171,66
S/P	99,66%	99,30%	99,82%	99,65%	99,65%	99,20%

TABLE 1 – Récapitulatif des résultats des mitigations du biais.

Suppression totale : des scénarios de suppression de variables ont été construits en utilisant le niveau de dépendance entre le genre et les autres variables explicatives. La performance et l'équité ont été mesurées pour chaque scénario puis un des scénarios non-dominés a été choisi. Ce choix est discutable par rapport au niveau de perte de performances que l'acteur est prêt à consentir dans le but d'obtenir un modèle plus équitable. Quatre variables ont été supprimées pour permettre d'obtenir l'arbitrage présenté, certaines de ces variables étaient significatives pour la modélisation du risque. Les variables restantes ont pu dans une certaine mesure remplacer ces variables. Ainsi, le biais a pu être réduit tout en maintenant un niveau de performances satisfaisant.

∗ *Apports :* consiste en une sélection assez simple des variables. Le travail effectué en amont pour définir et calculer le coefficient HGR permet d'obtenir des niveaux de dépendances exploitables. Les mesures de dépendances classiques captent des dépendances trop faibles pour permettre des distinctions entre les variables.
∗ *Limites :* la réussite de cette approche dépend de la capacité des variables non supprimées à compenser la perte d'informations induite par la suppression des variables.

Suppression des corrélations : en faisant varier l'hyperparamètre permettant de moduler le niveau de suppression, 100 scénarios ont été construits. Parmi ces scénarios, des scénarios dominant le modèle de référence ont été trouvés, ils sont plus équitables et plus performants.

∗ *Apports :* un scénario dominant le modèle de référence a été trouvé. L'approche est simple à implémenter car basée sur l'utilisation d'une régression linéaire.
∗ *Limites :* cette méthode est contraignante car pour qu'elle fonctionne, il faut des variables explicatives quantitatives et idéalement une variable sensible quantitative. Et même si c'est le cas, il faut, après correction, discrétiser les sorties et pouvoir faire des correspondances aux moments de la prédiction des primes.

Adaptation fair-SMOTE : cette méthode a permis de rééquilibrer efficacement les distributions des primes par genre tout en conservant la forme initiale des distributions. Comme le montre les résultats, la perte de performance est négligeable. Toutefois, dans le cas de la tarification, cela ne suffit pas pour réduire significativement le biais historiquement présent dans les données.
L'algorithme de génération des individus, les hypothèses de rééchantillonnage, la discrétisation des primes ont été remis en question pour tenter d'obtenir de meilleurs résultats,

sans succès.

∗ *Apports :* l'augmentation des données a été de bonne qualité, les différents scénarios et hyperparamètres définis permettent d'adapter l'algorithme aux besoins. Cette méthode pourrait être plus utile dans des cas où la représentation des genres dans les données est un problème central.
∗ *Inconvénients :* un effet relativement négligeable sur le biais historique et sur les inter-dépendances.

Exponentiated gradient : cette méthode est la seule ayant pu être implémentée comme approche de traitement pendant la modélisation. La contrainte intégrée est une contrainte d'égalité des erreurs de prédictions entre les genres. Bien qu'il soit théoriquement possible d'implémenter de meilleures conditions d'équité, ces contraintes n'ont pu être implémentées ni mathématiquement, ni algorithmiquement. Il se pourrait qu'elles ne puissent pas être appliquées telles quelles avec cette méthode. Faute de temps, cette piste n'a pas pu être explorée plus longtemps pour soit trouver une solution soit prouver l'existence ou non d'une solution.

∗ *Apports :* offre un cadre dans lequel des contraintes d'équité peuvent être directement ajoutées aux modèles.
∗ *Limites :* les temps de calculs sont exponentiels même pour une contrainte relativement simple. A cela s'ajoute la difficulté d'implémentation de contraintes adaptées au problème traité.

Redistribution équitable : En partant de l'équilibre initial dans lequel la prime pure contient une part significative de biais mesurée par l'adaptation du flip-test, un nouvel équilibre contenant moins de biais est obtenu en redistribuant les écarts de primes entre les genres. A cause des grandes dimensionnalités dans lesquelles sont représentés les individus, il n'est pas possible de corriger directement le biais en faisant : $prime_{equitable}\ i = prime_{biaisee}\ i - ecart_{flip-test}\ i$. En effet, cela a conduit à une amplification des biais car en rapprochant trop brusquement la prime d'un individu de la prime moyenne de ses voisins de genre opposé, elle s'éloigne potentiellement des primes d'autres individus dont il composait le voisinage. Des paramètres permettant de contrôler la vitesse de correction et le niveau de biais tolérés ont donc été introduits et hyperparamétrés. Le HGR KDE n'étant pas cohérent pour mesurer les effets de la redistribution, d'autres mesures ont été utilisées. Cette méthode a permis de réduire significativement les écarts entre les genres. Avant la redistribution, les hommes payaient en moyenne 1,3€ de plus que les femmes pour un total de 32019€. Après redistribution, cet écart passe à 2245€ au total et 0,13€ en moyenne.

∗ *Apports :* traite de l'équité individuelle dans le cas de la régression et adaptable à la problématique traitée.
∗ *Limites :* en plus de la qualité du modèle de prime, la qualité du modèle de k plus

proches voisins doit être surveillée. De plus, la distribution de cette solution n'est pas aisée. Il faut soit construire une grille avec toutes les corrections possibles soit construire un modèle distribuable permettant de prédire les écarts finaux.

Conclusion

L'équité est l'une des contraintes importantes à prendre en compte dans la tarification. Que celle ci soit imposée par la règlementation ou par les décisions stratégiques de l'entreprise, il faut être capable de définir, mesurer et mitiger le biais dans le but d'obtenir des modèles plus équitables. Ces besoins ont conduit en premier lieu à une étude de l'équité du point de vue mathématique. Les différentes mesures introduites ont permis de détecter du biais avant et après la modélisation quel que soit le traitement classique effectué sur le genre. Il s'en est donc suivi une étude des méthodes de mitigation. Ces mitigations ont été effectuées avant, pendant et après la modélisation avec pour objectif de traiter le biais au global en retraitant les données ou en imposant des contraintes d'équité, mais aussi de traiter le biais de manière individuelle en retraitant les différentes primes. Certaines de ces méthodes ont été plus concluantes que d'autres mais des leçons ont pu être tirées de chaque méthode.

Dans la continuité de cette étude, plusieurs jeux de données et garanties pourront être étudiés et ainsi obtenir des résultats plus fiables et un benchmark de référence. Il faudra également envisager d'autres mailles de tarification telles que la maille portefeuille ou groupe de garanties.

Ces travaux posent un cadre pour la mesure et la mitigation des biais en assurance, un cadre dans lequel de nombreuses pistes peuvent être explorées. Par exemple, de meilleures méthodes de mitigation pendant la modélisation peuvent être implémentées et le traitement simultanée de plusieurs variables sensibles peut être étudié.

Executive summary

Context

Insurance is characterized by the inversion of its production cycle : the insurer asks for a fixed premium at the subscription in exchange for coverage against risks whose occurrence and amount are uncertain. This inversion emphasizes the statistical nature that surrounds pricing. This must be consistent with statistical theory to allow the estimation and coverage of random phenomena. Beyond these statistical considerations, premiums represent the fair price of the insurance service, and like any price, it encapsulates many strategic challenges. Insurance is an increasingly competitive market. The latter brings together both major historical players, but also new, more agile players who are trying to recover market share. All of this is stimulated by constantly changing regulations that try to promote competition between players. Thus, a player in the insurance market must offer competitive prices, aligned with the company's strategy and communication, but also use tools adapted to the different distribution processes.

The last few years have been marked by the use of so-called machine learning algorithms and neural networks as well as massive data. This is due to scientific advances but especially to the increase in computing power, its accessibility and the large amount of data in circulation. These new technologies are impacting several business sectors and insurance is no exception. These data and algorithms serve as decision-making support tools, they help segment the insured, understand the risk and all the different factors that surround it. The actors must therefore integrate these new elements in order not to lose their competitiveness. However, the use of this massive data and complex algorithms quickly put the subject of transparency on the table ; premiums and decisions must be explainable and fair. Indeed, the stakes are high, it is not about algorithms helping to choose a film but about algorithm setting the cost of access to insurance services for segments of the population. These issues of fairness and ethics have been an integral part of our societies and thoughts for centuries. Although subject to interpretation, fairness can be defined as the ability to put individuals on an equal footing by taking into account the differences that exist between them. It is out of the principle of fairness that rules such as the gender directive have emerged. Reforms on access to borrower insurance are also a form of fairness enforcement. Thus, players in the insurance sector may have to build fairer premiums in relation to so-called sensitive or protected variables. These fairness

constraints can arise from regulations, as is the case with gender, or from commercial and strategic objectives, as is the case with age in certain companies.

In these studies, the concept of bias dealt with is bias in the sense of discrimination, it designates the undesirable effect of a sensitive variable on a variable of interest, for example the effect of gender on the premium in the framework imposed by the gender directive. Since 2016, many research studies have pointed out the presence of discrimination in decision-making tools : the tool for measuring the risk of recidivism in the United States, the algorithms of Google images, Amazon application processing, lending and financing algorithms and many more. For a long time, the solution used has been to circumvent the problem without solving it from an ethical point of view. A simplified situation in which a gender-correlated variable V is used to determine the premium has been staged. Gender is not taken into account in the estimates, it is fairness by omission. However, the presence of correlation between gender and the insensitive explanatory variable leads to an indirect effect of gender on the estimated premiums. Thus, although gender is apparently reprocessed or removed from the model, its relationship with the other variables allows its effect on the target variable to be retained. This is due to the fact that the restatements only make it possible to deal with the direct effect of the sensitive variables and not their indirect effects, whereas these variables exert a non-negligible influence on the distributions of the other explanatory variables. Age, just like gender or disability, affects activity choices, risk taking, product choices, etc. It is therefore essential to first be able to define and measure the fairness of the models constructed.

Definition and measurement of fairness

The literature on the study of bias in models is quite recent. References and consensus are not explicit as described by American researchers in April 2021 : "The rapid growth of this new field has led to extremely inconsistent motivations, terminologies and notations, presenting a serious challenge for the cataloging and comparison of definitions". In the 12th edition of the March 2022 Scientific Report on data science and artificial intelligence, the authors talk about the zoo of definitions of fairness : "the researcher or practitioner approaching this facet of machine learning for the first time can easily feel confused and somehow lost in this zoo of definitions.These multiple definitions capture different aspects of the concept of fairness but, at the best of our knowledge, there is still no clear understanding of the overall landscape where these measures live."

These studies offer a review of the literature on the subject, an almost exclusively English-language literature. The review will not be exhaustive but clear and comprehensive enough to be used to address various insurance fairness issues. In addition, the case of classification is preponderant in the literature, these studies therefore propose methods and adaptations to allow fairness to be taken into account in the continuous case imposed by pricing.

Fairness is understood intuitively as the absence of dependence between the sensitive variable and the variables of interest. Indeed, independence implies that no direct or indirect relationship exists between these variables and bias cannot be present. This definition based on independence is only one aspect of fairness. It is first necessary to distinguish the framework of observations from the framework integrating causality. The causal framework, although interesting, is not treated in detail because the statistical tools for establishing causality are still too limited. Then, it is necessary to discern group fairness and individual fairness. Group fairness or statistical fairness stipulates that certain statistics must be the same between the individuals of the different groups formed by the sensitive variable, this is the classic approach. Individual fairness, on the other hand, specifies that similar individuals should be treated in the same way. This approach is the most intuitive but also the most difficult and costly to implement because it is linked to causality.

Finally, several metrics have been defined in an attempt to quantify the level of fairness. Six measures are retained for the use-case, classical measures such as the mean ratio and Kendall's tau and measures of differences between probability distributions such as the Kolmogorov-Smirnov test and the JS divergence. In addition to these measures, two measures occupy a central place in this study, these are the HGR KDE and the adaptation of the flip-test.

HGR KDE : this metric is theoretically the strongest. It makes it possible to measure all types of dependencies between all types of variables. Two versions have been implemented ; the version based on neural networks and the version using kernels (KDE). The KDE version was chosen because it converges twice as fast as the one based on neural networks. Moreover, the choice of the optimal structure of the neural network is delicate. A lighter structure would allow faster convergence but possibly less reliable results.

Adaptation flip-test : this metric makes it possible to define individual fairness without requiring the construction of a graph and a causal model. This is possible under the assumption that a distance or an algorithm can measure the proximity between the individuals in the database. Initially built for a binary target variable, it was adapted in this work using a k-nearest-neighbor algorithm as a measure of proximity. Thus, it is possible to compare the prediction of an individual with that of the closest individuals of the opposite gender. The quality of the notion of proximity provided by the model of k nearest neighbors built is the main limit of this approach. To control this limit, hyperparameterization and variable selection are carried out. The chosen distance and its setting as well as the number of neighbors are optimized. The number of variables and the variables retained to build the model are also optimized, the objective being to obtain the model of k nearest neighbors leading to the smallest distance between individuals but also the lowest premium differences. Thus, on these different parameters, a search by grid is carried out and the results are checked on a test database. This approach provides a satisfactory solution with a significantly lower operational cost than that which a causal

study would generate.

After detecting the bias, it is essential to attempt to mitigate it while preserving the performance and consistency of the models constructed.

Bias mitigation

Bias mitigation is the act of taking action on the models, data or results in order to reduce the bias between the sensitive variable and the estimate of the variable of interest. As with measurements, there is no consensus on approaches to mitigating bias. The field of research is expanding and the approaches proposed in the literature are in some cases too closely related to the use case they deal with. The binary Y case is here again the preferred case. In this work, the methods that can be used for the case of regression are presented. Mitigation can be practiced before, during or after modeling.

Ante modeling : these methods amount to transforming the data in order to reduce the bias while preserving the relevant information. The methods implemented are the total removal of variables correlated to the sensitive variable, the removal of linear correlations and an adaptation of the fair-SMOTE method. Initially built for the binary case, this method is adapted to meet the needs of mitigation in non-life pricing.

∗ *Advantages :* these methods have the advantage of keeping the entire modeling process intact. They are generally simple to implement and less costly in computation time.
∗ *Disadvantages :* these methods can lead to significant loss of information. The variables may be closely related to each other and not allow both the removal of bias and the retention of information. In addition, it will be necessary to ensure that the rest of the modeling process does not introduce bias.

During modeling : these methods consist of integrating fairness constraints directly into the model calibration phase. Two methods are implemented, the exponentiated gradient which takes its name from the game theory method on which it is based, and its grid search version.

∗ *Advantages :* these methods use information on the sensitive variable in order to find the best balance between performance and fairness. Theoretically, they are more likely to lead to the best possible trade off.
∗ *Disadvantages :* these methods are difficult to implement and generalize. Moreover, even after a successful implementation, their convergence is not guaranteed and the computation times can be exponential.

Post-modelling : it consists of transforming the predictions obtained at the output of the models in order to make them fairer. The first example that is easy to think of is that of logistic regression. The probabilities obtained make it possible to modify the decision

thresholds with the consequence of impacting the behavior of the model with respect to the different classes. The post-modeling methods present in the literature are therefore methods for modifying decision boundaries taking into account equity measures. These approaches are not applicable to the continuous case because there are no decision boundaries. Furthermore, no publication applying and testing post-modeling mitigation in this case could be found.

Studies have therefore been carried out with the aim of proposing a mitigation approach in the continuous case. In order to be able to impose equity after the calculation of the premiums, it would be necessary to be able to define what is an advantageous premium and what is a disadvantageous premium. However, this is not possible in absolute terms because an advantageous premium depends on the characteristics of the insured and the risk carried. It is following this observation that the individual "equitable redistribution" approach was proposed. It exploits the adaptation of the flip-test as a definition of the bias to segment the premium into a fair share and a bias. This bias is then redistributed to the individuals in the base in an iterative manner, with the aim of converging towards a state in which, given the new premiums, the measured bias is as low as possible.

∗ *Advantages :* these methods do not require recalibrating the models, so they have shorter computation times. The outputs obtained can no longer be contaminated by bias because they are the final results.

∗ *Disadvantages :* the results are dependent on the quality of the models built, the mitigation on bad predictions is not very consistent.

Application to car insurance pricing

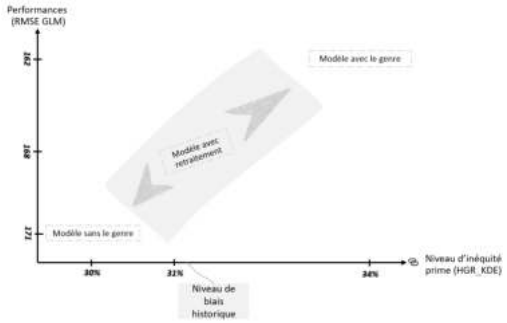
For this application, a glass breakage guarantee is priced taking into account, throughout the pricing process, statistical constraints but also business and operational constraints. The data was processed and analyzed, the variables reprocessed and discretized in order to build a modeling database. A database describing the vehicles was also integrated and a zoning variable was built. The RMSE and the total charge to premiums ratio, denoted S/P, were used, among other metrics, to evaluate the models built using GLM, random forest and gradient boosting. These models were optimized and then validated using interpretability methods.

GLM models were retained following this pricing phase. Indeed, although black box models can achieve better performance, GLMs remain the most widespread because they are simple to interpret and integrate into production pricing tools. Thus, the slight increase in performance obtained using the random forest model is not enough to select it because of the operational costs that this would generate. However, at each stage, the calculations are performed on all three models, and in general there are no significant deviations. Moreover, the pure premium model is preferred to a multiplicative cost $\times$ frequency model because it is more efficient in the sense of the defined metrics. Reference

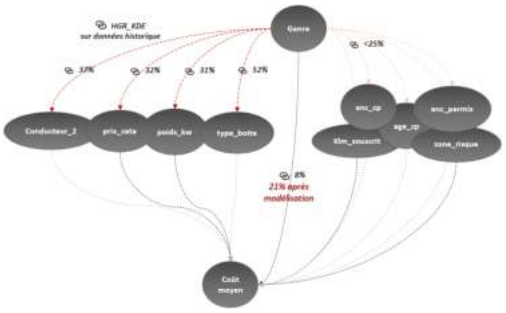
models are constructed by omitting gender. Once constructed, these studies go further by attempting to measure the gender bias and then mitigating it.

Bias measurement

Using the six dependence measures selected for this practical case, the biases are measured on the historical data and on the predictions. For simplicity reasons, the results on pure premiums are presented first in this note, the trends observed being the same for average costs and frequencies. Figure 2a summarizes the measurement results for the premium model.



(a) Trade-off between performance and fairness for the premium model.



(b) Direct and indirect effects of gender on average cost.

FIGURE 2 – Some results on the measurement of biases.

The various measures detected bias in the historical data, however, given that there are no reference measures to compare them with, it is difficult to draw thresholds and define the level of significance of these measurements.

A first model with gender is then built, it leads to higher levels of bias for each of the variables of interest. The bias in the cost model is multiplied by 2.6 and goes from 8% to 21%. Thus, the bias was not only learned by the model but it was also amplified. The analysis revealed that gender occupies an important role in the measurements of variables importance. In addition to this, the gender variable is linked to the other significant variables of the different models. The idea is that these interdependencies with gender have led to the amplification of its effect on the variables of interest.

The bias is then measured on a second model that does not contain the gender. The bias levels reached are close to the levels obtained in the historical data. Dependency measures such as Kendall's tau which detected very little dependency before modeling detects 2 to 3 times more after modeling. Thus, despite the absence of gender, it still exerts an influence on the model's predictions. As in the simplified example, deleting gender removes its direct effect but does not deal with its indirect effect through the other explanatory variables. Figure **??** illustrates these different effects for the average

cost.
In multivariate models such as those constructed here, these gender distinctions may arise from combinations of several variables. For example, a combination of vehicle power, box type, and risk zone can be very unbalanced in terms of male-female distribution. For example, by crossing zoning and vehicle prices, significant imbalances appear between genders in the dataset. 87% of individuals with a vehicle costing more than $30,000€$ and living in the zone 12 are men. Thus, any significant difference in premium between this segment and the others has a strong impact on men.

In order to verify that these levels of bias are not the result of a poor consideration of gender in the initial pricing, a model reprocessing gender at the output of the models was built. It leads to substantially the same results, as the reprocessing does not solve the problem caused by interdependence.
Tests showed that the explanatory variables could predict gender by using classification models. It therefore appears that the variables can reconstruct gender and therefore retain its effect on predictions. In order to obtain more fair models, it is essential to apply more fair-sensitive mitigation methods.

Bias mitigation

The various mitigation approaches are tested taking into account the trade-off between performance and fairness. Depending on the trade-off that the actor decides to make, the best models can be modified. Mitigation is performed directly on the pure premium model. The notion of non-dominated model is defined in order to allow the elimination of inconclusive scenarios. A model is said to be non-dominated when there is no other model having both better performance and better fairness than it. Thus, according to the trade-off agreed by the insurer, a non-dominated model may be the best and not be according to another trade-off. The gender-neutral model is used as a benchmark for both performance and fairness. For each of the methods, sensitivities to the choice of hyperparameters and modeling choices are put in place. The results of these mitigations are compiled in the table 4.22. They come from the five methods whose results, contributions and limitations are presented below.

Metrics	Reference model	Total deletion	Correlation removal	Adaptation fair-SMOTE	Exponentiated gradient	Equitable Redistribution
HGR KDE	29,71%	19%	26,50%	28,83%	31,22%	30,08%
RMSE	171,04	177,6	169,6	171,61	171,25	171,66
S/P	99,66%	99,30%	99,82%	99,65%	99,65%	99,20%

TABLE 2 – Summary of bias mitigation results.

Total deletion : variable deletion scenarios were constructed using the level of dependence between gender and the other explanatory variables. Performance and fairness

were measured for each scenario and then one of the non-dominated scenarios was chosen. This choice is questionable in relation to the level of loss of performance that the actor is ready to accept in order to obtain a more fair model. Four variables were deleted to obtain the trade-off presented, some of these variables were significant for the risk modeling. The remaining variables could to some extent replace these variables. Thus, the bias could be reduced while maintaining a satisfactory level of performance.

∗ *Contributions :* consists of a fairly simple selection of variables. The work done upstream to define and calculate the HGR coefficient makes it possible to obtain exploitable levels of dependencies. Classical dependency measures capture dependencies that are too weak to allow distinctions between variables.
∗ *Limits :* The success of this approach depends on the ability of the undeleted variables to compensate for the loss of information caused by the deletion of the variables.

Correlations removal : by varying the hyperparameter allowing to modulate the level of suppression, 100 scenarios were built. Among these scenarios, scenarios dominating the reference model were found, they are more fair and more efficient.

∗ *Contributions :* a scenario dominating the reference model has been found. The approach is simple to implement because it is based on the use of linear regression.
∗ *Limits :* this method is restrictive because for it to work, quantitative explanatory variables and ideally a quantitative sensitive variable are required. And even if this is the case, it is necessary, after correction, to discretize the outputs and to be able to make correspondences when premiums predictions will be needed.

Fair-SMOTE adaptation : this method has effectively rebalanced the distributions of bonuses by gender while maintaining the initial shape of the distributions. As the results show, the performance loss is negligible. However, in the case of pricing, this is not enough to significantly reduce the bias historically present in the data.
The algorithm for generating individuals, the resampling hypothesis, the discretization of premiums have been questioned in an attempt to obtain better results, without success.

∗ *Contributions :* the data augmentation was of good quality, the different scenarios and hyperparameters defined allowed the algorithm to be adapted to the needs of the study. This method might be more useful in cases where gender representation in the data is a central issue.
∗ *Limits :* a relatively negligible effect on historical bias and interdependencies.

Exponentiated gradient : this method is the only one that could be implemented as a processing approach during modeling. The built-in constraint is a prediction error equality constraint between genders. Although it is theoretically possible to implement better fairness conditions, these constraints could not be implemented either mathematically or algorithmically. They may not be able to be applied as is with this method.

Due to lack of time, this track could not be explored any longer to either find a solution or prove the existence or not of a solution.

∗ *Contributions :* provides a framework in which fairness constraints can be directly added to models.
∗ *Limits :* computation times are exponential even for a relatively simple constraint. Added to this is the difficulty of implementing constraints adapted to the problem at hand.

Equitable redistribution : starting from the initial equilibrium in which the pure premium contains a significant portion of bias measured by the adaptation of the flip-test, a new equilibrium containing less bias is obtained by redistributing the differences in premiums between genders. Because of the large dimensionality in which individuals are represented, it is not possible to correct the bias directly by doing : $premium_{fair}\ i = premium_{unfair}\ i - gap_{flip-test}\ i$. Indeed, this led to an amplification of the biases because by bringing the premium of an individual too abruptly closer to the average premium of his neighbors of the opposite gender, it potentially moves away from the premiums of other individuals whose neighborhood he made up . Parameters making it possible to control the speed of correction and the level of tolerated bias have therefore been introduced and hyperparameterized. Since the HGR KDE is not consistent in measuring the effects of redistribution, other measures have been used. This method significantly reduced gender gaps. Before the redistribution, men paid on average 1,3€ more than women for a total of 32019€. After redistribution, this gap decreases to 2245€ in total and 0.13€ on average.

∗ *Contributions :* deals with individual fairness in the case of regression and can be adapted to the addressed problem.
∗ *Limits :* In addition to the quality of the premium model, the quality of the k-nearest-neighbor model should be monitored. Moreover, the distribution of this solution is not easy. It is necessary either to build a grid with all the possible corrections or to build a distributable model making it possible to predict the final deviations.

Conclusion

Fairness an important constraints to consider in pricing. Whether it is imposed by regulation or by the company's strategic decisions, it is necessary to be able to define, measure and mitigate the bias in order to obtain fairer models. These needs led in the first place to a study of fairness from a mathematical point of view. The various measures introduced detected bias before and after the modeling regardless of the classic treatment carried out on gender. A study of mitigation methods therefore followed. These mitigations were performed before, during and after the modeling with the aim of treating the bias globally by reprocessing the data or by imposing fairness constraints, but also to

treat the bias individually by reprocessing the different premiums. Some of these methods were more successful than others, but lessons could be learned from each method.

In the continuity of this study, several sets of data and guarantees can be studied and thus obtain more reliable results and a reference benchmark. It will also be necessary to consider other pricing meshes such as the portfolio mesh or group of guarantees.

This work nevertheless provides a framework for measuring and mitigating biases in insurance, a framework in which many avenues can be explored. For example, better mitigation methods during modeling can be implemented and the simultaneous treatment of several sensitive variables can be studied.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'assurance a une place importante dans les sociétés modernes. Elle permet de favoriser le développement économique tout en protégeant les populations des retombées financières parfois graves des risques qui entourent leurs quotidiens. Ce rôle social place l'assurance dans une position particulière. Comme toutes entreprises, elle doit faire face à la concurrence, être rentable pour les actionnaires, les sociétaires etc. mais elle doit aussi respecter une réglementation stricte, être transparente et équitable.

La tarification, étant l'un des éléments centraux de l'assurance, n'échappe pas à ces différentes contraintes opérationnelles, commerciales et règlementaires. Cette tarification en assurance non-vie requiert une segmentation des assurés en se basant sur leurs informations, sur le bien à assurer et sur toutes autres sources de données externes cohérentes. Il s'agit donc d'une forme de discrimination basée sur les informations de l'assuré. La question de l'équité est donc primordiale car la limite entre discrimination et segmentation est fine.

Le genre en tarification automobile est l'exemple le plus connu de la mise en place de l'équité en assurance. En effet, depuis 2012, la gender directive oblige les assureurs à mettre fin à toute discrimination basée sur le genre des assurés. Ainsi, les assureurs ne doivent plus présenter de différences de primes entre la gent féminine et la gent masculine. Face à cette contrainte, les deux solutions les plus répandues sur le marché sont :

- Le retrait de la variable genre de tous les modèles de tarification. Cette approche s'appelle « l'équité par omission » ;
- La prise en compte de la répartition homme/femme au sein d'un portefeuille pour, in fine, présenter une prime unique.

En apparence ces méthodes permettent d'atteindre l'équité car les primes présentées ne sont pas distinguées par genre. En réalité, l'équité atteinte n'est que fictive car les interrelations existantes au sein des grands volumes de données à disposition de l'assureur peuvent introduire de façon indirecte l'influence de la variable genre dans les modèles. Des variables pourtant non sensibles comme la puissance de la voiture, son modèle, les zones et les offres souscrites peuvent être utilisées par les modèles pour reproduire la présence du genre. Cette influence indésirable du genre sur les primes, du fait de l'équité recherchée, est appelée biais dans ces travaux. Ce biais n'est pas à confondre avec les biais statistiques qui eux sont bien connus et traités. Il s'agit plutôt d'un biais éthique,

d'un biais de discrimination. Face à cette nouvelle problématique, il est donc nécessaire d'avoir les outils spécifiques permettant de détecter, de mesurer et d'inhiber ce biais pour ajuster les modèles actuariels et de science des données et ainsi assurer l'équité tout en maintenant la qualité et la performance des modèles.

Afin de répondre à cette problématique en fournissant les outils théoriques et opérationnels nécessaires, ces travaux s'articulent en quatre grandes parties. D'abord, l'assurance et ses différents enjeux sont présentés ainsi que la tarification et ses différentes contraintes. Cette partie permet de poser le décor et de se familiariser avec les différentes notions qui joueront un rôle important dans les décisions de modélisation. S'en suivent deux parties qui présentent le cadre théorique et les outils mathématiques permettant la mesure et la mitigation du biais. Enfin, une dernière partie qui, pour un cas réel de tarification non-vie, mesure et tente de mitiger le biais tout en préservant la qualité et la cohérence de la tarification.

L'équité en apprentissage statistique est un domaine assez récent ; c'est à partir de 2016 que l'intérêt pour ce sujet est suscité et que de nombreux travaux sont produits. Et comme tout nouveau domaine, il requiert du temps pour atteindre des consensus, une harmonisation des approches et des définitions, et des références. Les travaux sont, pour la plupart, issus de la littérature anglophone et traite en majorité du cas de la classification, le cas le plus intuitif pour introduire, mesurer et traiter le biais.
Ces sujets ont aussi attiré l'attention de chercheurs de la communauté actuarielle tels que Mario Wüthrich, Arthur Charpentier, Marcin Detyniecki et leurs différents co-auteurs. Ainsi, début 2022, deux travaux de recherches tentant de définir l'équité et un cadre dans lequel la mitigation des biais peut être appliquée, en utilisant par exemple des réseaux neuronaux, sont publiés.

L'apport des travaux réalisés ici est de s'appuyer sur une revue extensive de la littérature pour proposer une présentation claire et complète des notions d'équité. Cette présentation ne sera, certes, pas exhaustive mais elle permettra de donner une vision globale et précise du sujet pour ces travaux et ceux à venir. De plus, contrairement à la majorité de la littérature, ces travaux s'étendent sur le cas de la régression, cas plus complexe mais indispensable pour la tarification non-vie. Ainsi, de nombreuses approches et adaptations s'inspirant de la littérature sur la classification sont proposées. L'objectif étant de tenter de fournir une référence à la littérature actuarielle francophone permettant de disposer d'abord des outils mathématiques puis opérationnels nécessaires à la mise en place de l'équité.
En juillet 2022, Arthur Charpentier dans OPINIONS & DÉBATS de l'institut Bachelier publie un article de qualité vulgarisant les questions d'équité en assurance de manière générale, preuve de l'intérêt grandissant pour ce sujet dans les sciences actuarielles.

Chapitre 1

La tarification non-vie : entre rigueur statistique, stratégies et enjeux sociétaux

Dans cette partie, il s'agit de présenter l'importance de l'assurance dans la société et la place de la tarification. S'en suit une brève présentation des étapes de tarification avant des discussions sur les contraintes et la transparence de ces processus.

1.1 Assurance et tarification

Le risque est un facteur omniprésent de la vie de tous les humains. Il peut prendre la forme d'un accident de la circulation, d'un incendie, d'une maladie, d'un décès, d'un accident de travail etc. Ces risques sont une source significative d'incertitudes et leurs potentielles retombées financières expliquent la nécessité de l'assurance. L'assurance est née du besoin des humains de se couvrir contre l'incertitude qui entoure leurs activités en s'entraidant. Pour un individu pris tout seul, cet aspect social de l'assurance est quasi invisible. Les personnes s'assurent à cause du caractère obligatoire de certaines assurances (la responsabilité civile en automobile) ou parce qu'elles sont risquophobes et préfèrent payer une prime relativement faible plutôt que de faire face à des retombées plus grandes. Personne ne s'assure en se disant : "ma prime permettra de couvrir mes éventuels sinistres, ceux des autres et de favoriser le développement de l'économie..."

Cette vision sociale est pourtant très importante dans une assurance qui fonctionne en grande partie sur les principes de la mutualisation et de la diversification des risques. Dans une note rédigée par Denis Kessler, Amélie de Montchatlin et Christian Thimann, il est discuté que l'assurance favorise la croissance économique, la stabilisation et l'innovation dans l'économie[47].

Croissance et développement économique. Les bienfaits de l'assurance sur le développement économique sont palpables. Néanmoins, mesurer de façon objective son im-

pact est difficile. Cela s'explique notamment par le fait que le développement économique engendre le développement de l'assurance qui lui favorise celui de l'économie. De plus, l'impact de l'assurance peut être assimilé en certains points à celui du secteur financier ; le développement de l'assurance ne peut se faire qu'en présence des institutions financières nécessaires à la distribution de ses services. En dépit de ces limites, une étude menée par la Banque Mondiale (Arena, 2006) a montré que l'assurance, et surtout l'assurance non-vie, a un effet positif significatif sur le développement de tous les pays quels que soient les niveaux de revenus.

Stabilisation. L'assurance permet d'aider les individus à faire face à des chocs collectifs comme les catastrophes naturelles ou les crises économiques. Elle permet aussi aux individus de faire face à des chocs personnels : un grave accident de la circulation ou des frais de soins élevés. Les aléas de la vie peuvent rapidement faire passer un individu d'une situation financière confortable à une situation de quasi-faillite. L'assurance agit ici comme un filet de secours.

Distribution / Mutualisation. En mutualisant les risques des assurés, l'assureur crée une forme de solidarité entre eux. Des primes de quelques dizaines d'euros prises une à une permettent de régler des sinistres de plusieurs milliers d'euros. Ainsi, l'assureur met en place une forme de distribution entre ses assurés.

Innovation. En fournissant un matelas de sécurité, l'assurance favorise l'innovation. En effet, elle protège des chocs externes, préserve les patrimoines et permet une prise de risque plus maîtrisée.

Les assureurs ont donc un rôle social important. Ce rôle, ils le jouent en échange des primes collectées. Leurs calculs requièrent une segmentation des assurés qui s'effectue suivant les informations à la disposition de l'assureur, c'est-à-dire des informations sur l'assuré, sur le bien à assurer mais également toutes autres sources de données externes cohérentes qui pourraient être utilisées pour quantifier le risque. Avec ces informations, l'assureur forme des sous-groupes homogènes d'assurés dans lesquels il pourra appliquer un certain niveau de prime. Une prime est donc une forme de "discrimination" basée sur une représentation du risque porté par l'assuré.
Ces primes sont le prix du service d'assurance, et comme tout prix, il encapsule de nombreux enjeux stratégiques. En effet, l'assurance est un marché de plus en plus concurrentiel. Ce dernier regroupant à la fois de grands acteurs historiques, mais aussi de nouveaux acteurs plus agiles qui tentent de récupérer des parts de marché. Le tout étant stimulé par une réglementation en constante évolution qui tente de favoriser la concurrence entre les acteurs. Ainsi, un acteur du marché de l'assurance doit proposer des prix compétitifs. Cette grande pression sur le niveau des primes met dans certains cas les considérations statistiques au second plan.
Toutefois, en plus de ces considérations commerciales et statistiques, les primes sont d'une importance cruciale car elles définissent le coût de l'accès aux services d'assurance pour

les différents segments de la population. Le rôle de l'assurance dans la société est donc une contrainte importante à prendre en compte dans la constitution de la tarification finale.
Ainsi, pour des raisons réglementaires ou sociales, certaines discriminations ne sont pas admises dans les primes. Par exemple, l'assureur automobile ne peut pas afficher des primes différentes par genre alors que cette variable est souvent significative dans la segmentation du risque.

Dans ces travaux, il sera discuté que le traitement d'effet de variables non souhaitées nécessite un travail approfondi et que la simple suppression de la variable explicative dans la définition de la prime d'assurance n'est pas suffisante. Au-delà de ce volet social et réglementaire, un assureur peut s'imposer des contraintes supplémentaires en accord avec sa stratégie de développement. L'assureur peut vouloir, pour des raisons commerciales, présenter des primes qui ne contiennent pas de discriminations entre certaines parties de la population. Quels que soient l'origine et l'objectif d'une contrainte, elle peut se résumer à rendre un modèle équitable par rapport à une variable. L'assureur doit donc être équipé d'outils lui permettant d'ajuster ses primes sans détruire leur cohérence et leur compétitivité.

L'assurance est caractérisée par l'inversion de son cycle de production, l'assureur demande une prime fixée à la souscription en échange d'une couverture contre des risques dont la réalisation et le montant sont aléatoires. Cette inversion met l'accent sur le caractère statistique qui entoure la tarification, des études doivent être menées dans le but d'estimer le coût des garanties qu'il propose. Ces modélisations introduisent par définition des biais statistiques qu'il est nécessaire de distinguer du biais éthique traité dans ces travaux.

1.1.1 L'approche de tarification classique

La tarification en assurance non-vie est un processus long pouvant nécessiter des études sur une période allant de 3 à 9 mois. Dans ce processus, il est utile de prendre en compte l'étude ou la conception de l'offre et l'étude des conditions des contrats avant toute modélisation. Une fois la modélisation terminée, de nombreux ajustements sont appliqués sur les primes et les modèles obtenus. Des études sont aussi menées dans le but de mesurer les impacts de la nouvelle tarification.

En assurance non vie, la tarification au sens de modélisation statistique des risques peut se décomposer en les étapes suivantes :

1. Analyser les besoins et objectifs commerciaux.
2. Constituer les bases de données qui serviront à la tarification.
3. Analyser les données et les traiter.
4. Construire les variables de modélisation.
5. Modéliser les variables d'intérêt, la fréquence, la sévérité ou directement la prime pure.
6. Évaluer, améliorer et valider le modèle construit. Pour cela plusieurs allers-retours sont effectués entre les phase d'analyse, de traitement et de modélisation.

7. Enrichissement de la tarification, par la construction de zonier, la prise en compte d'interaction ou l'utilisation de véhiculier.
8. Modélisation finale. À la suite de ces travaux, le modèle définitif est construit et ajusté si nécessaire pour répondre aux besoins et objectifs commerciaux.

Ces différents éléments sont abordés en détail dans le chapitre 4 au moment de la tarification. A chaque étape de ce processus, un équilibre est recherché entre les enjeux de l'entreprise et la rigueur statistique. Il faut être rigoureux dans le but d'approcher le plus fidèlement le risque et de pouvoir faire face à ses engagements mais en même temps il faut être capable de faire preuve de pragmatisme pour obtenir des tarifs cohérents avec les objectifs de l'entreprise. La construction de la tarification n'est donc pas un exercice technique dissocié de toute logique de marché. Au delà de la recherche de l'équilibre statistique, de nombreuses contraintes pèsent sur ce processus.

1.1.2 Les contraintes de tarification

Les deux principales contraintes prises en compte au moment de la tarification sont les contraintes opérationnelles/commerciales et les contraintes réglementaires.

Les contraintes opérationnelles et commerciales. Il est assez fréquent que l'assureur modifie la prime obtenue pour des raisons commerciales : se repositionner sur le marché, gagner des parts de marché etc. Certains assureurs s'attaquent à des segments spécifiques du marché en baissant leurs primes. D'autres encore modifient leurs primes pour être en accord avec la politique et l'image qu'ils voudraient transmettre au marché. Par exemple, certains assureurs prônent la solidarité générationnelle ce qui voudrait dire que malgré la distinction entre les risques par âge ceux-ci ne seront pas pleinement pris en compte dans la tarification.

Le processus de tarification est sujet à de nombreuses décisions de modélisation. Par exemple, les variables à discrétiser, les méthodes à utiliser, les modalités à regrouper, les périmètres d'études etc. doivent être choisis. Tout ces choix sont faits en accord avec la théorie statistique et les contraintes commerciales. De plus, l'assureur prend en compte les coûts de la mise en place des solutions implémentées. Ainsi, avant toute modification de son processus de tarification, l'assureur s'interroge sur la viabilité du système à mettre en place, sa rapidité, sa facilité de prise en main et son coût. Par exemple, bien que statistiquement l'utilisation de modèles boîtes noires accompagnés de méthode d'interprétabilité peut sembler intéressant, l'assureur est souvent confronté à une incompatibilité de tout ou partie de sa chaîne de tarification en aval de la modélisation et doit alors ajouter dans la balance les différents coûts et risques afférents à toute évolution. En plus de ces contraintes opérationnelles ou commerciales, l'assureur doit aussi évoluer dans un environnement réglementé.

Les contraintes réglementaires. L'assureur a l'obligation d'expliquer la prime obtenue pour chaque assuré. Sa tarification doit être transparente et juste. En plus de

devoir expliquer la prime, l'assureur doit respecter des règles anti-discriminations. En effet, du fait de la gender directive, l'assureur n'a plus le droit depuis le 21 décembre 2012 de proposer une prime différente entre les hommes et les femmes. La directive Anti-discrimination[19] vise à faire interdire l'utilisation de variables tarifaires jugées discriminantes telles que l'âge ou la présence de handicap. L'utilisation massive des données personnelles devraient entraîner des évolutions de la réglementation dans les années à venir.

Bien que ces variables jugées discriminantes permettent d'améliorer la modélisation des risques, l'assureur se retrouve contraint de ne pas les utiliser dans les primes à proposer.

Ainsi, quelles que soient les raisons pour lesquelles un assureur doit réduire voire annuler l'effet d'une variable dans ses modélisations, il doit être équipé d'outils lui permettant de le faire tout en conservant la performance et la cohérence de ces modèles.

Le niveau de règlementation de l'assurance est un bon baromètre de sa place dans la société. La règlementation permet de disposer d'un meilleur accès à l'assurance et d'une diversité de choix, d'avoir un respect des engagements des assureurs et un contrôle des différentes pratiques. Les questions anti-discriminations mettent au premier plan les questions d'éthique et d'équité en assurance. Il y a des années, les autorités se sont demandées si l'utilisation du genre était acceptable dans la prime proposée. Plus récemment, dans l'accès à l'assurance emprunteur, la question s'est posée de savoir s'il fallait pénaliser ou non les personnes ayant des maladies graves comme le cancer. Ces questions sont des questions d'ordre social permettant d'aboutir à la prise de résolutions et ainsi améliorer l'inclusion ou l'accès aux services pour des groupes spécifiques d'assurés.

1.1.3 Introduction à la notion d'équité

L'équité est d'abord une notion étroitement liée à la philosophie et au droit. Ses débuts remontent aux écrits d'Aristote et à la loi romaine. Elle est définie comme le fait de traiter justement tous les individus en respectant ce qui leur est dû. André Comte-Sponville[17] la définit comme étant "la vertu qui permet d'appliquer la généralité de la loi à la singularité des situations concrètes et qui vise à instaurer une égalité de droit, en tenant compte des inégalités de fait". En d'autres termes, l'équité revient à pouvoir appliquer l'égalité en tenant compte des différences qui existent entre les individus.

L'équité renvoie donc à l'éthique dans le sens où elle prône que les individus doivent être traités dans le but d'égaliser leurs différentes chances par rapport à une situation donnée. Tandis que l'égalité renvoie plus à la morale en souhaitant que les individus soient traités de la même manière quelles que soient leurs caractéristiques. La figure 1.1 provenant de l'encyclopédie canadienne offre une illustration des différences entre équité et égalité.
Ainsi, la mise en place de l'équité peut mener en certains points à traiter les individus concernés différemment. En assurance non-vie, une tarification équitable reviendrait à présenter des primes ne contenant aucune discrimination contre certaines tranches de la société, ces tranches étant définies implicitement par ce que la société considère comme éthique. Ainsi, des variables comme le genre, l'âge, les origines, l'orientation sexuelle,

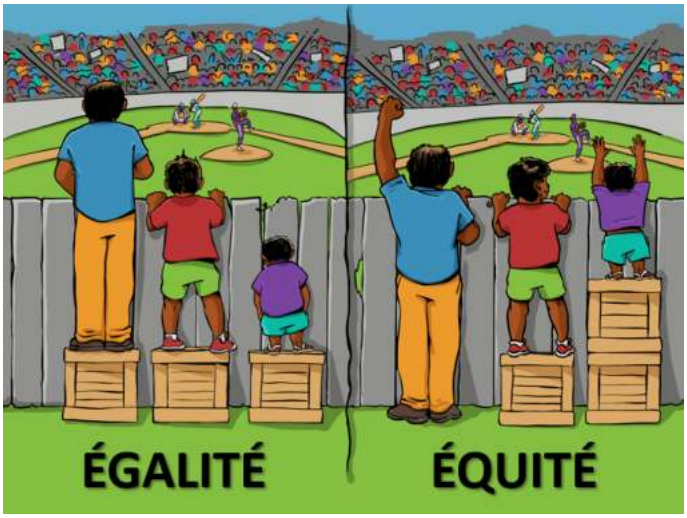


FIGURE 1.1 – Différences entre égalité et équité.

la situation familiale, la présence de maladies graves ou héréditaires et d'handicap sont jugées sensibles. L'équité pourrait être vue comme le fait d'avoir des niveaux de primes égaux quelles que soient les valeurs de la variable sensible. Pour la maladie par exemple, cela voudrait dire que la présence de maladie ne sera pas un facteur différenciant pour la détermination des primes. Cette définition de l'équité introduit implicitement la notion de discrimination positive dans le sens où un individu malade porte plus de risques qu'un individu non malade. Par conséquent, les primes équitables seront vraisemblablement plus fortes pour l'individu non malade qu'elles ne devraient l'être en réalité. Il apparaît donc que l'équité soit en quelques points contradictoire avec la tarification fidèle du risque.

Un sujet très proche de l'équité est celui de la transparence de la tarification. Il est impossible de parler d'équité pour des modèles non transparents et vice versa. Ce sujet a connu un essor ces dernières décennies du fait du développement de la grande famille de l'intelligence artificielle. De plus grands volumes de données, plus de puissance de calcul, des méthodes ayant des structures de paramètres gigantesques ont envoyé la notion de boîte noire dans une nouvelle dimension. La transparence et l'interprétabilité ont tout de suite tapé à la porte pour rappeler que dans certains cas, le gain de performance ne pouvait justifier la perte de transparence.

1.2 Interprétabilité et transparence

L'intelligence artificielle regroupe l'ensemble des méthodes et techniques utilisées dans le but de permettre aux ordinateurs de simuler les capacités intellectuelles humaines. A l'intérieur de cette famille, se trouve l'apprentissage statistique. Ce sont les méthodes fondées sur la théorie mathématiques et statistiques qui permettent d'estimer des fonctions de prédiction à l'aide de données.

Les avancées dans la recherche couplées avec l'augmentation significative de la puissance de calcul et de l'accès à cette puissance ont conduits à une plus grande utilisation de ces méthodes. Ces nouvelles méthodes (nouvelle dans le sens de leur application) ouvrent de nombreuses possibilités aux assureurs. Elles permettent d'automatiser des tâches, d'améliorer les processus clients, de mieux comprendre le client, de détecter les fraudes etc. Il y a de grands enjeux sur le plan compétitif. En effet, les acteurs du marché qui ne sauront pas profiter de ces avancées technologiques, risquent de perdre en compétitivité et de mettre en danger leur part de marché.
L'utilisation de modèles de plus en plus complexes et des données massives a mis en exergue les questions de transparence des modèles. En effet, l'augmentation du nombre de variables utilisées ou l'utilisation de modèles complexes dit "boîtes noires" rendent les résultats fournis par les modèles peu lisibles pour les utilisateurs. Cette limite n'en est pas toujours une : dans certains cas, un modèle ne nécessite pas d'être interprété. Par exemple, si les erreurs du modèle n'engendrent aucune conséquence grave (recommandation de musique sur youtube), ou dans les cas où la justesse du modèle a été étudiée et prouvée empiriquement (reconnaissance de texte). Dans ces cas, le coût de l'interprétabilité est trop grand par rapport au gain de lisibilité.

Par contre, en assurance, cette illisibilité est inadmissible. L'assuré et l'assureur doivent être capable de connaître les facteurs qui conduisent à une certaine prime. Cette contrainte provient de la place qu'occupe l'assurance dans la société. Plusieurs assurances sont imposées par la loi ou sont quasi obligatoires pour avoir accès à certains services ou biens. De plus, l'assurance reste un marché fortement concurrentiel et sans transparence de la part de l'assureur, l'assuré se dirigera facilement vers un autre assureur. D'un autre côté, cette capacité à comprendre les algorithmes produisant les primes apporte de nombreux avantages développés plus bas. Entre autres, elle facilite la détection et de la mitigation des biais pour viser l'équité au sein d'une population donnée. Cette dernière étant un trait permettant de définir un algorithme transparent.

1.2.1 Utilité de la transparence des modèles

Il existe de nombreuses raisons de vouloir mieux comprendre le comportement des modèles, les quatre principales sont les suivantes :

L'interprétabilité est une exigence réglementaire. La présence de systèmes de prise de décision automatique nécessite un cadre réglementaire adapté. En effet, ces systèmes sont manipulables, peuvent contenir des biais et être mal construits. Il est donc nécessaire d'être sûr qu'ils ne créent pas de préjudices sociaux. Ainsi, le Règlement Général sur la Protection des Données, RGPD, dans son article 22 encadre les prises de décisions automatisées lorsqu'elles affectent les individus de manière significative ou conduisent à des effets juridiques. Par exemple, dans le cas du refus d'une demande de prêt bancaire, le demandeur peut demander à savoir pourquoi le système refuse sa demande.

Confiance et vérification. Un modèle pour lequel il n'est pas possible de vérifier ou d'expliquer ses décisions ne peut remplir pleinement son rôle. En effet, dans plusieurs cas le "pourquoi" est plus important que la prédiction elle-même. Il permet d'orienter les utilisateurs des modèles et vérifier que le modèle se base sur des phénomènes cohérents pour sa prise de décision. La transparence fournie par l'interprétabilité donne cette impression de participation à la prise de décision. Ce volet humain est précieux. Il facilitera l'adoption de l'intelligence artificielle (IA). Tant que cette confiance ne sera bâtie, son utilisation se limitera à des cas à très faible impact (choisir un film, une vidéo, filtrer des mails, etc...).

Nouvelle connaissance. L'intelligence artificielle a la capacité d'extraire des données des informations qu'il serait normalement impossible d'extraire. Le fait de pouvoir interpréter les décisions des modèles permettrait d'avoir accès à ces nouvelles connaissances, découvrir de nouvelles interactions entre les variables dans des conditions spécifiques, découvrir des relations de cause à effet, etc.

Détecter les biais. Les systèmes d'IA souffrent des limites de leurs qualités, ils sont capables de représenter les données dans de grandes dimensions pour en extraire un maximum d'informations. Cela leur permet de découvrir des relations cachées mais aussi cela conduit à l'apprentissage de biais difficiles à détecter. Ils sont capables de détecter et d'amplifier tous les biais présents dans les données.

Les modèles sont aussi capables de se construire des circuits de prédictions sans réellement résoudre le problème. Dans la littérature, l'histoire du Cheval Hans[65] est assez bien documenté. Il s'agit d'un cheval qui savait apparemment compter dans les années 1900. En réalité, le cheval avait appris à détecter dans les gestes de son dresseur des indices qui lui permettaient de trouver les bonnes réponses. Le dresseur n'avait aucune idée que le cheval se servait de son langage corporel pour trouver les réponses. Les modèles récents de détection ont souvent eu de tels comportements en identifiant les arrières plans des images plutôt que les objets eux-mêmes[52, 61, 48].
L'interprétabilité des modèles permet de détecter ces biais qu'ils soient issus des données ou des modèles eux-mêmes. Comprendre comment fonctionne un modèle est indispensable pour pouvoir l'améliorer et le contrôler. Ce sujet est depuis quelques temps l'objet de nombreux travaux. L'interprétabilité devra obligatoirement devenir une étape inévitable de l'entraînement des modèles.

La transparence a aussi fait l'objet de travaux dans le secteur de l'assurance. L'Autorité de Contrôle Prudentiel et de Résolution (ACPR) recommande[20] l'évaluation des modèles et outils utilisant l'intelligence artificielle suivant les 4 principes d'explicabilité, de performance, de stabilité et de traitement adéquat des données.

1.2.2 Explicabilité

La définition d'un modèle interprétable dans un cadre général peut présenter des difficultés du fait du manque de consensus mathématiques. Il existe néanmoins des travaux qui vont dans le sens de poser un cadre universel de discussion. La limite entre ce qui est compréhensible et ce qui ne l'est pas est parfois fine. Elle peut dépendre de celui qui regarde : ce qui est compréhensible pour un peut ne pas l'être pour un autre. Certains utilisateurs ont besoin d'avoir accès à plus de détails que d'autres. Aussi, suivant le type de problème, certaines interprétations n'ont plus de sens. Dans le cadre de la tarification non-vie, un modèle sera jugé interprétable de façon satisfaisante quand il sera possible d'attribuer à chaque variable un effet sur le niveau de la prime. Cette définition qui représente le résultat recherché en pratique colle bien avec la définition de Miller de 2017 : "L'interprétabilité est le degré auquel un humain peut comprendre la cause d'une décision." Certains modèles comme les modèles linéaires généralisés ou les arbres de décisions sont interprétables de par leur construction. Tandis que d'autres comme les forêts aléatoires ne permettent pas directement d'avoir accès aux éléments ayant entraîné une prédiction. En plus d'expliquer les prédictions individuelles, les concepteurs des modèles auront intérêt à comprendre le comportement global du modèle et à pouvoir extraire les éléments clés sur lesquels s'appuient les modèles pour leur apprentissage. L'explicabilité n'est pas qu'une exigence réglementaire. En forçant la compréhension des modèles et de leurs sorties, cela permet la correction et l'ajustement des résultats.

Les enjeux de performances, de stabilité et de traitement des données sont eux aussi d'une grande importance ; même s'ils ne sont pas au cœur de ce mémoire.
Le traitement des données permet d'obtenir des données de qualité qui permettront d'obtenir un bon modèle car un modèle mal nourrit donnera de mauvais résultats. Il permettra aussi de s'assurer aux premiers abords que les données utilisées pour la construction des modèles respectent toutes les exigences métiers et réglementaires.
La performance met en avant le fait que les modèles doivent remplir leur rôle. Ils doivent produire des estimations satisfaisantes du phénomène qu'ils permettent de modéliser. Les modèles doivent aussi être performants du point de vue opérationnel (i.e temps de réponses, facilité d'accès etc.) et du point de vue normatif (i.e prévention contre les fraudes, répondre à toutes les exigences normatives etc.).
La stabilité met l'accent sur le fait que les performances du modèle doivent être maintenues dans le temps. Les modèles doivent pouvoir être entretenus dans le but de faire face à toutes modifications de l'environnement dans lequel ils ont été entraînés. Ces modifications peuvent être réglementaires, techniques ou économiques.

Au-delà de l'explicabilité, la littérature présente d'autres principes importants permettant de définir la transparence, l'une d'entre elles est l'équité.

Équité comme notion de transparence. Un algorithme équitable est un algorithme dont les résultats ne conduisent à aucun biais ou effet discriminant envers une catégorie particulière de la population. La discrimination peut être introduite consciemment ou inconsciemment par ses concepteurs. La discrimination peut aussi provenir des données

utilisées pour la modélisation. En assurance, ce sujet est subtil. En effet, la tarification est une forme de discrimination par le risque. Il est néanmoins important de s'assurer que des biais indésirables ne sont pas pris en compte.

1.2.3 Le coût de la transparence

La transparence a un coût. Les modèles nativement interprétables sont généralement des modèles moins précis et plus restrictifs tandis que les modèles plus complexes sont des boîtes noires difficiles à cerner. La première solution envisagée dans la littérature est de trouver des méthodes pour interpréter les modèles plus complexes. Ce sont en général les méthodes agnostiques au modèle définies en annexe. Ces méthodes demandent beaucoup de temps de calculs et ont leur lots d'imperfections, du fait d'être manipulables à la non prise en compte de multicolinéarités et d'interactions par exemple. La seconde approche est de construire des modèles complexes mais interprétables. Cette approche conduit à des modèles moins précis que les modèles boîtes noires classiques. La transparence se paye donc au prix de la précision, du temps de calcul ou de l'exhaustivité des explications.

Sur le marché de l'assurance, il y a de nombreux facteurs à prendre en compte, les enjeux sont divers et parfois opposés. Il faut pouvoir trouver un équilibre satisfaisant pour les assurés et les acteurs du marché dans le respect de l'éthique et de la règlementation. La figure 1.2 montre quelques éléments impactant les décisions prises dans le monde de l'assurance.

FIGURE 1.2 – L'assurance, un marché subissant de nombreuses pressions.

Dans ces travaux les enjeux concernant l'équité des modèles sont étudiées. Le volet social

apparaît lorsque les variables sensibles sont dictées par des perceptions sociales et par la règlementation. A contrario, le volet stratégique est plus présent lorsque l'assureur décide de prendre les devants et de réduire l'effet d'une variable sur sa tarification. La rigueur statistique est de mise tout au long de ces analyses aussi bien que les contraintes classiques de tarification, l'objectif étant d'étudier l'équité dans un cadre réaliste.
Depuis assez longtemps l'approche de traitement utilisée est d'ignorer le problème ou de le contourner sans réellement le résoudre. Les variables sensibles étant simplement omises ou les résultats retraités après modélisation. Ces approches ne sont toutefois pas suffisantes. Pour être en mesure de rendre un modèle équitable par rapport à une variable donnée, il faut être capable de définir, de détecter et de mesurer les biais causés par cette variable pour ensuite pouvoir réduire ce biais en appliquant des approches conscientes/sensibles aux enjeux d'équité.

Chapitre 2

Définition, détection et mesure du biais dans un problème d'apprentissage statistique

2.1 Machine learning et biais : cadre théorique

Initialement, la modélisation se faisait en tirant des idées et des conjonctures de l'observation de phénomènes physiques ou biologiques. Les modèles étaient construits pour permettre de répliquer le comportement des phénomènes sous des hypothèses définies par leurs auteurs. Et quand de nouvelles observations rendaient difficile la vérification de certaines hypothèses, des extensions étaient construites pour relâcher les hypothèses des modèles initiaux. À l'ère du big data, le paradigme est différent. La grande quantité de données disponibles est supposée suffisante pour décrire le monde réel. Pour décrire et prédire la réalité, il faut donc ajuster un modèle efficace à ces données. La tarification non-vie n'échappe pas à ce nouveau cadre. En effet, aucun modèle mathématique n'a été construit dans le but de modéliser le comportement de la sinistralité. Les modèles de tarification sont plutôt construits sur l'hypothèse que les observations faites sur l'historique de sinistralité permettent en grande partie de définir la réalité du risque. Il s'agit de l'apprentissage statistique supervisé. L'objectif est de construire une fonction prédictive capable d'apprendre les relations entre les variables explicatives X et la variable cible Y. Cette fonction permettra d'utiliser des réalisations x pour prédire des y inconnues. Pour construire cette fonction, n couples de variables $(X_1, Y_1), ..., (X_n, Y_n)$ sont disponibles. Ces variables sont supposées indépendantes et identiquement distribuées (i.i.d.) issues d'une loi de probabilité inconnue $\mathbb{P}$. Seule la loi empirique est connue à travers les observations disponibles dans le jeu de données.

Le modèle sera évalué à l'aide d'une fonction de perte pour une prédiction $\widehat{y}$ avec y connue, $L : (y, \widehat{y}) \mapsto L(y, \widehat{y}) \in \mathbb{R}^+$. Avec cette fonction de perte, la construction du modèle devient un problème de minimisation. Ainsi, trouver la meilleure fonction f dans

un ensemble donné de fonctions $\mathbb{F}$ (algorithmes) revient à la minimisation suivante :

$$\arg\min_{f\in\mathbb{F}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(X_i)) \right\}.$$

Une pénalisation peut être rajoutée pour éviter le sur apprentissage :

$$\arg\min_{f\in\mathbb{F}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(X_i)) + \lambda p(f) \right\},$$

où p est une fonction de pénalisation et λ permet de réguler le niveau auquel est pénalisé le modèle.

L'un des principaux biais qui entoure ce processus de modélisation est l'erreur de généralisation. Cette erreur vient du fait que le modèle ne se soit pas performant sur les données réelles. Cela peut être dû à la qualité des données et/ou des modèles.
Au niveau des données, l'hypothèse suivant laquelle les observations sont issues de la même loi de probabilité est importante. Elle justifie le fait de construire un modèle sur des observations pour ensuite le généraliser à la population. Si cette hypothèse est significativement fausse, le modèle ne pourra pas se généraliser quelle que soit sa qualité sur les données disponibles.
Le modèle, de son côté, peut souffrir de biais de modélisation. Ce biais est le fait que la fonction prédictive construite se trompe systématiquement dans l'estimation de Y. Cela peut s'expliquer par le fait que la classe de fonction prédictive choisie ne soit pas adaptée pour capter toutes les relations présentes dans les données. Par exemple, utiliser un modèle linéaire lorsque des relations non linéaires significatives existent dans la base de données. Cette erreur est le fruit d'un sous-apprentissage. En apprentissage statistique, la décomposition biais-variance fournit un cadre d'analyse des erreurs de prédiction d'un modèle. Elle stipule que l'erreur d'un modèle se décompose en trois éléments : le biais lié au sous-apprentissage, la variance liée au sur apprentissage et l'erreur irréductible.

- Biais : Biais $= \mathbb{E}[\widehat{f}] - f$. Le fait que les prédictions du modèle dévient en moyenne des valeurs réelles.
- Variance : Var $= \mathbb{E}[(\mathbb{E}[\widehat{f}] - \widehat{f})^2]$. Le fait que le modèle fasse du sur-apprentissage en apprenant les bruits présents dans les données.
- l'erreur irréductible ϵ. Cette erreur peut soit provenir de l'aléa intrinsèque au phénomène modélisé, soit de la non-exhaustivité des données pouvant être mises à disposition.

La figure 2.1 permet de visualiser les concepts de biais et variance.

Un modèle trop simple aura un biais élevé mais une variance faible. En complexifiant le modèle dans le but de réduire son biais, sa variance devient de plus en plus grande. Il s'agit du "trade off biais-variance". Il n'est pas possible de minimiser ces deux valeurs en même temps. La figure 2.2 offre une visualisation de cet échange entre biais et variance. Les deux figures précédentes proviennent d'une série de publication wikipedia sur les

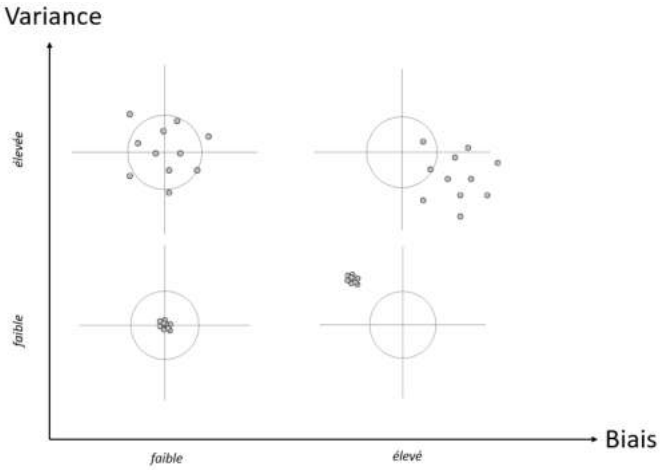


FIGURE 2.1 – Illustration des concepts biais et variance.

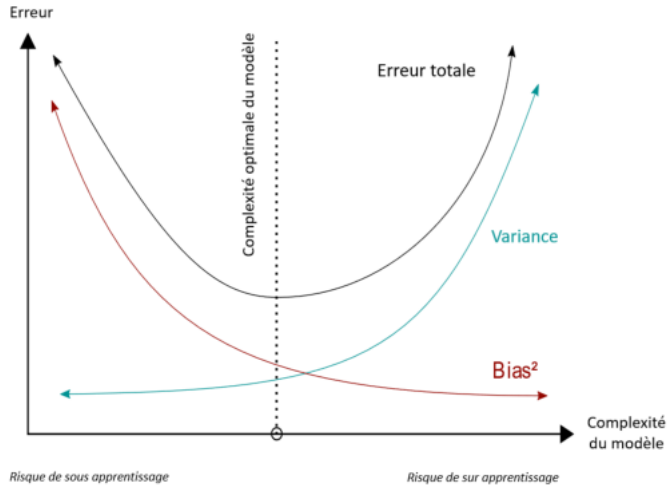


FIGURE 2.2 – Trade off entre biais et variance arbitré par la complexité du modèle.

erreurs de modélisation[79].

Ces différents biais sont assez bien connus et traités dans la théorie statistique. La validation croisée permet d'avoir une vision plus robuste de la capacité de généralisation du modèle. La pénalisation et les bons hyperparamétrages permettent de contenir le sur-apprentissage.

La notion de biais étudiée dans ces travaux est différente de celles évoquées plus haut. Dans ce mémoire, il est question d'étudier le biais dans le sens de discrimination. Traiter le biais dans ces travaux revient à rendre un modèle équitable par rapport à des critères imposés en amont de la modélisation. Ainsi, le cadre est celui d'un acteur qui pour des raisons réglementaire ou stratégique souhaite traiter les individus de classes différentes

(homme/femme pour le genre) de manière similaire par rapport à une variable cible (par exemple la prime). Au-delà de l'équité, les notions classiques de qualité statistique des modèles et d'adéquation commerciale ne devront pas être négligées.

2.2 La discrimination en tarification non-vie

En assurance, la segmentation joue un rôle important dans la détermination des primes. L'idée de l'assurance est de rassembler et de mutualiser le plus grand nombre de risques. En rassemblant des risques qui se ressemblent, l'assureur, avec l'appui de la loi des grands nombres, devrait en moyenne pouvoir couvrir ces risques. Il lui suffirait d'appliquer comme tarification pour tous la moyenne de la charge de sinistre. Une prime ainsi calculée ferait référence à une mutualisation totale du risque.

Toutefois, la présence d'anti-sélection et de concurrence rend l'utilisation d'une mutualisation totale incohérente. En effet, en mutualisant tous ses risques, l'assureur prend le risque de regrouper des profils de risque significativement différents. Ainsi, les bons risques auront une prime trop élevée du fait de la présence de mauvais risques. Ils se tourneront donc vers la concurrence. Les mauvais risques quant à eux auront des primes plus faibles par rapport aux risques qu'ils portent. Ces mauvais risques continueront donc d'entrer dans le portefeuille. La mutualisation sera mise à mal. L'assureur est donc dans l'obligation de segmenter ses risques ; construire des sous-ensembles de risques homogènes où il pourra appliquer la mutualisation sans subir d'anti-sélection. Il utilise donc tous les facteurs de risques à sa disposition dans le but de constituer des sous groupes d'assurés et de calculer des primes pour chaque sous groupe. L'assureur effectue donc une discrimination entre les assurés à l'aide de leurs informations personnelles, des informations sur leurs biens et des données externes dites "open data". Il ne peut cependant pas discriminer à l'extrême puisque la loi des grands nombres ne pourrait plus s'appliquer. Les primes et les résultats de l'assureur deviendraient trop volatiles.

En plus de faire attention à la sur-segmentation, l'assureur accorde une attention particulière à son positionnement par rapport à la concurrence. Cela le conduit à reconsidérer les primes ou segments obtenus d'un point de vue métier/commercial. Ainsi, en plus des contraintes légales et éthiques, le positionnement commercial peut être un des éléments qui pourrait pousser un assureur à réduire les effets de certaines variables sur ces modèles. L'étude de la détection et de la mitigation du biais est donc indispensable pour permettre aux acteurs d'avoir des modèles à la fois significativement équitables et pertinents en tout point.

La détection et la mitigation de biais non souhaités en assurance sont toutefois délicates. Le but initial étant justement de construire des modèles exploitant au mieux les différences entre les individus. L'approche classique de mitigation du biais est de négliger les variables sensibles dans la construction des modèles (équité par ignorance ou omission). Toutefois, cette approche ne permet de traiter que la discrimination directe et néglige l'effet de la discrimination indirecte.

Discrimination directe vs discrimination indirecte. Pour illustrer les limites de l'équité par omission un exemple simplifié est étudié ci-dessous. A la suite de cet exemple les notions de discrimination directe et indirecte sont présentées. Dans cet exemple, les tableaux montrent respectivement les nombres de sinistres et l'exposition pour un contrat automobile.

$n_{i,j}$	Femme	Homme	$n_{i,\bullet}$
A	45	8	53
B	20	53	73
$n_{\bullet,j}$	65	61	126

$e_{i,j}$	Femme	Homme	$e_{i,\bullet}$
A	151	33	184
B	129	290	419
$e_{\bullet,j}$	280	323	603

TABLE 2.1 – Exemple 1 : Le nombre $n_{i,j}$ et l'exposition $e_{i,j}$

Ces informations sont données dans un tableau croisé entre 2 variables : la variable genre et une variable explicative V ayant deux classes A et B. En considérant la variable genre comme étant une variable non utilisable pour la tarification du risque, son effet sur les primes devient indésirable, il devient un biais. L'approche classique pour le traitement de ce biais est de négliger cette variable dans les calculs. Ainsi, pour cet exemple simplifié, en prenant en compte nombre de sinistres et expositions, les estimateurs suivants sont obtenus :

$$\widehat{\mu}_{1,\bullet} = \frac{n_{1,\bullet}}{e_{1,\bullet}} = \frac{53}{184} = 0.288 \tag{2.1}$$

$$\widehat{\mu}_{2,\bullet} = \frac{n_{2,\bullet}}{e_{2,\bullet}} = \frac{73}{419} = 0.174$$

Donc, une distinction est à première vue effectuée uniquement entre les individus des classes A et B sans la prise en compte du genre (les agrégats étant calculés tous genres confondus). Les individus de la classe A auront des primes plus élevées que ceux de la classe B. Il faut toutefois remarquer que les variables genre et V sont fortement corrélées. 82% des individus de la classe A sont des femmes. Cette corrélation conduit à une prise en compte indirecte du genre dans les calculs. En effet, la relation (2.1) peut se réécrire :

$$\widehat{\mu}_{1,\bullet} = \widehat{\mu}_{1,1}\widehat{\mathbb{P}}(\text{Femme}|A) + \widehat{\mu}_{1,0}\widehat{\mathbb{P}}(\text{Homme}|A) = \widehat{\mu}_{1,1}\frac{e_{1,1}}{e_{1,1}+e_{1,0}} + \widehat{\mu}_{1,0}\frac{e_{1,0}}{e_{1,1}+e_{1,0}} \tag{2.2}$$

Il apparaît donc que l'estimation obtenue dépend de la variable V mais aussi d'une manière indirecte, de la répartition des genres dans les classes. Ainsi, avec ces estimations, les femmes paieront plus cher que les hommes du fait de leur liaison avec la classe A. La suppression de la variable genre a fait disparaître son effet direct sur la prime mais indirectement, à travers V, le genre continue d'influencer les résultats. La figure 2.3 montre les relations qui existent entre les variables de cet exemple.

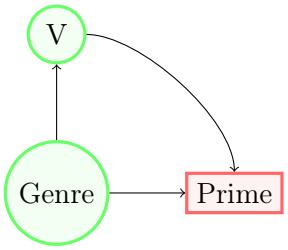


FIGURE 2.3 – Les relations de causalité présentent entre les variables

En assurance automobile, il est connu que la variable genre permet d'expliquer une partie de la sinistralité. Cette variable est étiquetée comme étant discriminatoire. Il s'agit d'une variable sur laquelle les assurés n'ont aucune influence. De plus, la question qui se pose est de savoir si la significativité de cette variable ne provient pas d'autres variables qui ne sont pas exploitées (la prise de risque au volant, la vitesse moyenne de conduite etc.). Ainsi, il est discriminatoire d'attribuer une prime à un individu sur la base des comportements des individus du même genre tout en sachant que le genre est en grande partie imposé à la naissance.

La variable V peut être une variable qui, prise individuellement, ne permet pas d'expliquer le risque mais qui du fait de sa relation avec S devient significative. Par exemple, la couleur de la voiture. La variable V peut d'autre part être une variable qui est à la fois significative pour expliquer la sinistralité et corrélée avec S. Par exemple la puissance de la voiture dans certains cas. La difficulté sera donc de réduire les effets indirects de S tout en préservant les capacités explicatives des autres variables. Dans la pratique, des relations inter-variables plus complexes que celles illustrées dans l'exemple peuvent permettent de conserver l'effet des variables sensibles.

Dans la suite des travaux, S dénote l'ensemble des variables sensibles ou protégées. Ce sont les variables interdites par la réglementation ou que l'acteur ne souhaite pas utiliser comme variables discriminantes. X les variables explicatives utilisables pour la modélisation et Y la variable cible. f dénote une fonction permettant de faire le lien entre les variables explicatives et la variable cible. Du point de vue de la performance, l'assureur aimerait utiliser les variables des ensembles S et X dans le but d'exploiter le maximum d'informations. L'estimateur de Y s'écrit donc :

$$f(X, S) = \mathbb{E}[Y|X, S].$$

En supposant que X, S et Y appartiennent au même espace de probabilité,

$$\mathbb{E}[\mathbb{E}[Y|X, S] - Y] = 0.$$

Cet estimateur est donc sans biais. Il est le meilleur que l'assureur puisse obtenir à l'aide des données à sa disposition. Toutefois, cet estimateur discrimine les individus suivant la variable protégée S. Cette discrimination est celle devant être mesurée et mitigée.

La prime obtenue par omission est la suivante :

$$f(X, S) = \mathbb{E}[Y|X].$$

Cet estimateur est sans biais dans les mêmes conditions énoncées précédemment. En cas d'indépendance entre X et S, l'omission de S permet de faire disparaître toute forme de discrimination. En pratique, il n'y a pas d'indépendance, la variable protégée ayant souvent un effet significatif sur le modèle et étant liée aux autres variables explicatives. Il est donc important de distinguer les discriminations directe et indirecte, l'omission, comme explicité dans l'exemple 1, ne permettant de traiter que la forme directe. Par définition la discrimination directe est traitée si :

$$f(Z) = \mathbb{E}[Y|Z],$$

avec Z une variable aléatoire $\sigma(X)$-mesurable. En posant $Z = X$ et en calculant les espérances sur les mêmes espaces probabilisés, il apparaît que l'omission empêche la discrimination directe.

De manière plus générale, une prime empêche la discrimination indirecte si :

$$Z \perp\!\!\!\perp S.$$

Il est clair que l'absence de discrimination indirecte entraîne l'absence de la discrimination directe. En effet, Z et S ne peuvent être indépendantes si Z contient des éléments de S. Il est important de garder en vue qu'en pratique, il n'est pas possible de rendre Z et S complètement indépendantes. De plus, en prenant en considération l'objectif de précision recherché dans les modèles, l'équité totale coûterait sûrement trop chère en terme de précision à l'acteur qui l'implémenterait. Avant de pouvoir mesurer le biais, il est important de comprendre son origine.

2.3 Les origines du biais

Dans cette section, la provenance des biais est étudiée. Les biais peuvent avoir plusieurs sources. Ces sources peuvent être rangées en 3 grands groupes :

— Biais des données vers l'algorithme ;
— Biais de l'algorithme vers l'utilisateur ;
— Biais de l'utilisateur vers les données.

Dans chacun de ces grands groupes peuvent être classé un grand nombre de sources de biais. Les deux principales étant le biais statistique et le biais historique.

2.3.1 Le biais statistique

Le biais statistique ou biais de représentation provient du fait que la population sélectionnée ne soit pas représentative de la population générale. Cela peut être dû au fait que les mesures pour certains pans de la population ne soient pas de bonne qualité.

Aussi, lorsque la population observée ne provient pas d'une sélection aléatoire de la population totale, le terme biais de sélection est utilisé. Par exemple, pour un produit d'assurance, la population étudiée est celle ayant pu avoir accès au produit. Ce biais peut aussi se traduire par le fait que les résultats obtenus pour certaines parties de la population ne soient pas représentatives. Par exemple, si certaines populations sont plus sinistrés et que cela s'explique en partie par le fait que les routes soient moins bien aménagées ou qu'il y ait moins d'investissement sur la prévention et la sensibilisation, ces individus sont présents dans les données, mais ils ne sont pas observés dans des situations permettant une comparaison équitable avec les autres individus de la base de données.

Dans la pratique, il est fréquent que les classes défavorables soient en minorité dans les données. Ainsi, même si les données sont représentatives et correctement mesurées, le cadre de la modélisation en présence de classes déséquilibrées peut introduire du biais. Les individus moins représentés sont moins bien connus, les erreurs commises par les modèles peuvent être négligées etc. Dans certains cas, ce déséquilibre est une caractéristique de la population étudiée. Il ne s'agit donc pas d'un biais de représentation au sens stricte du terme. Il est parfois considéré comme un biais lié à la modélisation plutôt qu'aux données. Quoi qu'il en soit ces déséquilibres sont des aspects importants du biais statistique. Ils se présentent sous les trois formes détaillées ci-après.

Des données déséquilibrées. Le problème de données déséquilibrées apparaît lorsqu'une classe d'individus est sous-représentée dans le jeu de données. Cela peut se matérialiser suivants les modalités exposées plus bas.

Dans le cas d'une classification :

— Déséquilibre sur Y : une classe est plus représentée dans les données. Par exemple, les données contiennent 95% d'individus n'ayant eu aucun sinistre.

— Déséquilibre sur S : une classe de la variable sensible est plus représentée dans les données.

— Déséquilibre sur $S|Y$: pour certains Y, des classes de S sont sous représentées.

Les tables 2.2, 2.3 et 2.4 offrent des illustrations de ces déséquilibres.

Y	n_i
0	52
1	448
Total	500

TABLE 2.2 – Déséquilibre sur Y

S	n_i
0	522
1	78
Total	600

TABLE 2.3 – Déséquilibre sur S

Y/S	0	1	Total
0	200	60	260
1	45	195	240
total	245	255	500

TABLE 2.4 – Déséquilibre sur $S|Y$

Dans le cas d'une régression :

— Déséquilibre sur Y : certains points de la distribution de Y contiennent significativement moins d'instances.

— Déséquilibre sur S : une classe de la variable sensible est plus représentée dans les données.

— Déséquilibre sur $S|Y$: dans les distributions de $S|Y$, certaines zones sont significativement moins populées que d'autres pour certaines classes.

Les figures 2.4 et 2.5 offrent des illustrations de ces déséquilibres.

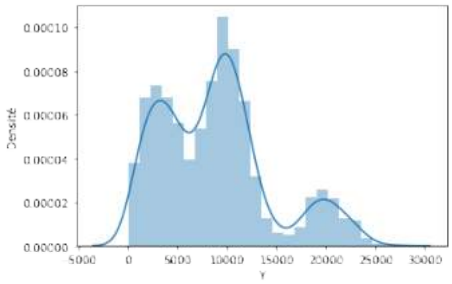


FIGURE 2.4 – Déséquilibre sur Y

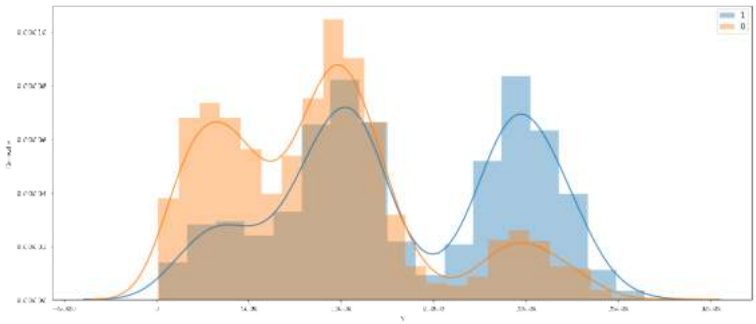


FIGURE 2.5 – Déséquilibre sur $S|Y$

Ce manque d'informations en certains points des distributions respectives peut conduire le modèle à commettre plus d'erreur en ces points ou tout simplement généraliser des comportements observés sur des échantillons non représentatifs.

Le biais statistique peut aussi apparaître du fait du paradigme de modélisation utilisé. En effet, l'apprentissage statistique est utilisé pour construire des fonctions prédictives basées sur les corrélations présentes dans les données. La causalité n'est pas intégrée dans la modélisation. Ainsi, dès le moment où les données utilisées ne sont pas exhaustives, des cofacteurs cachés peuvent introduire du biais. Imaginez un exemple dans lequel la sinistralité pour un produit d'assurance est corrélée avec le code postal alors que l'ethnicité est un facteur explicatif commun à ces deux variables (un cofacteur non observé). Le modèle basé sur les corrélations apprend indirectement à associer le risque à l'ethnicité.

2.3.2 Biais historique

Les données ont toujours une histoire. C'est pour cette raison que leur analyse demande du temps et des connaissances métiers. Ainsi, pour bien comprendre le biais présent dans la modélisation et ses origines, il faut comprendre le contexte qui entoure les données. Par exemple, en construisant un zonier en assurance automobile sur la France. Il apparaît souvent la "diagonale désertique". Un phénomène qui s'explique par le fait que historiquement, les zones situées sur cette diagonale soient moins peuplées. Cette connaissance du contexte permet de comprendre les résultats et de ne pas tirer des conclusions trop hâtives.
De plus, l'histoire est importante car les processus ayant conduit à la base définitive peuvent avoir introduit du biais dans les données. En effet, si à un moment la construction des bases de données a été sujette à des jugements ou décisions humaines, des biais peuvent avoir été introduits par les preneurs de décisions. Par exemple, en classant les bons et les mauvais risques, des biais conscients ou inconscients peuvent avoir été introduits dans les données par les équipes de sélections. Ces biais seront ensuite appris par les modèles, généralisés et se propageront dans le temps. Les données labellisées sont une illustration de comment les perceptions individuelles peuvent être amplifiées dans les modèles.

Des données mal labellisées. Les labels sont des éléments qui sont utilisés dans la phase de construction des bases de données dans le but de préciser leurs caractéristiques. Ces labels permettent ensuite d'entraîner des modèles de manière plus efficace.

Le processus de labellisation introduit du biais. Il y a le cas assez bien documenté datant de 2016 où la recherche "coupes de cheveux non professionnelles" présentait des coupes principalement de femmes noires[4, 41]. Tandis que la recherche "coupes de cheveux professionnelles" retournait des images de femmes blanches. La figure 2.6 montre les résultats qui étaient obtenus à l'époque sur google images.

FIGURE 2.6 – A gauche résultat pour "coupe de cheveux non professionnelle" et à droite "coupe de cheveux professionnelles" sur Google Images en 2016 ; Source : The Guardian

Les labels utilisés pour décrire ces images avaient conduit les algorithmes à faire ces classifications présentant des discriminations raciales. Ces labels provenaient d'articles écrits sur les coupes de cheveux, des préférences des différents auteurs. L'algorithme de Google a donc généralisé le biais qu'il a pu apprendre dans les données à sa disposition.

Biais historique : des faits statistiques ? Les biais historiques ont la particularité de devenir dans certains cas des faits statistiques. Ils ne sont plus considérés comme des biais mais sont considérés à tort comme des faits observés et justifiables. Le cas du genre est assez remarquable. Des décennies d'histoire ont créé des biais liés au genre. Ces biais sont si ancrés dans l'histoire qu'ils sont difficiles à dissocier du reste de la réalité. Il y a les exemples classiques des carrières professionnelles, de l'accès aux responsabilités et opportunités etc. Il y a un exemple plus subtil et intriguant. En 2018, une étude est menée pour savoir si l'écart de revenu entre les chauffeurs hommes et femmes sur la plateforme UBER est le fruit du sexisme des clients dans le choix de leurs courses. Le résultat de ces travaux est rassurant. La différence ne provient pas du sexisme mais plutôt des "différences entre les sexes dans les choix des conducteurs quant à l'endroit où conduire, la plus grande expérience des hommes sur la plateforme et la tendance des hommes à conduire plus vite." En prenant du recul sur ces résultats, et en les analysant avec le prisme du biais historique, ces différences ne sont-elles pas le fruit de décennies de sexisme ? Ces facteurs explicatifs ne sont-ils pas eux mêmes influencés par des biais provenant de la société ? La limite est donc très subtile entre réalité et discrimination. Il faut parfois prendre du recul et se demander si ces faits statistiques établis ne sont pas le fruit de discriminations indirectes.

2.3.3 Les autres sources de biais

Les notions suivantes sont issues des travaux de recherches sur les sources de biais en machine learning [62, 57, 75]. La figure 2.7 offre un aperçu des types de biais avec des exemples par type.

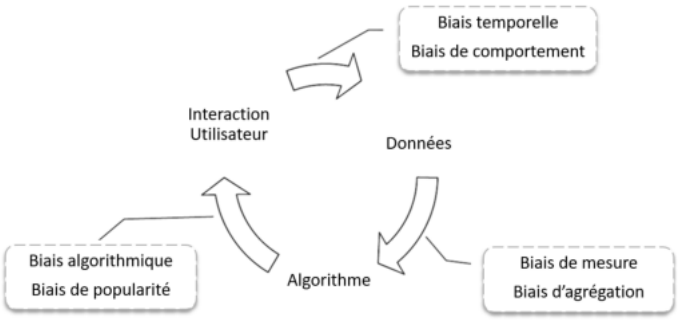


FIGURE 2.7 – Les différentes sources de biais

Biais de l'algorithme vers l'utilisateur. Les modèles construits à l'aide d'algorithmes sont des outils d'aide à la décision. Les résultats qu'ils fournissent modulent la perception qu'ont les utilisateurs du phénomène étudiée. Ainsi, un algorithme biaisé peut à son tour biaiser la perception des utilisateurs.

- **Le biais algorithmique** : il s'agit d'un biais introduit uniquement par l'algorithme utilisé. Les données ne sont pas biaisées. Cela peut provenir du mauvais choix des composantes de l'algorithme (pénalisation, structure, optimisation, poids des sous populations etc.).
- **Le biais de popularité** : ce biais apparaît du fait de la tendance des algorithmes à privilégier les éléments populaires. Ces éléments populaires sont ensuite plus valorisés par l'algorithme. La popularité n'est toutefois pas gage de qualité. Elle peut être nourrie par des facteurs externes biaisés.

D'autres biais de ce type existent : le biais émergent, le biais de l'interaction utilisateur et le biais d'évaluation.

Biais des données vers l'algorithme. Les données peuvent être biaisées. Comme expliqué plus haut, les modèles sont construits pour détecter et reproduire les éléments observés dans les données. Un jeu de données biaisé aura pour conséquence directe de biaiser tous les résultats. En plus du biais statistique, développé plus haut, les sources suivantes peuvent être citées :

- **Le biais de mesure** : ce biais apparaît lorsque les mesures ou les variables utilisées induisent des différences entre certains groupes d'individus. Pour que ceci soit considéré comme du biais, il faut que ces différences soient liées à des différences structurelles plutôt qu'au phénomène étudié.
- **Biais d'agrégation** : ce biais apparaît quand les études effectuées sur l'ensemble de la population conduisent à des conclusions erronées sur certains groupes d'individus.

D'autres biais de ce type existent : le biais d'omission et le biais de sélection.

Biais de l'utilisateur vers les données. La réalité est une notion relative. Ainsi, suivant les origines, l'éducation, les activités, la perception des réalités est différente. Les interactions entre utilisateur et données sont fréquentes. Certaines bases de données ont été construites à partir de données générées par des utilisateurs, d'autres ont été modelées par des utilisateurs. Par conséquent, les biais présents dans nos sociétés peuvent aisément se retranscrire dans les données qui seront utilisées pour les modélisations. En dehors du biais historique, les types de biais suivants peuvent être cités :

- **Biais de l'influence sociale** : ce biais apparaît lorsque l'avis ou le comportement des utilisateurs est influencé par le comportement des autres.
- **Biais temporel** : ce biais est lié au fait que des différences peuvent exister entre les populations dans le temps. Par exemple, en observant des données d'assurance sur les 10 dernières années, la période covid induit directement une différence dans les comportements des individus. Si ces différences ne sont pas prises en compte, elles peuvent introduire du biais dans la compréhension du phénomène.

— **Biais de comportement** : ce biais provient du fait que les individus n'ont pas le même comportement. Ainsi, suivant la population, le contexte, les outils, les contrats, le comportement des individus est différent.

D'autres biais de ce type existent : le biais de population, le biais d'auto-sélection et le biais de contenu.

Le biais discriminatoire ou éthique étudié est différents des biais statistiques usuels et provient de sources différentes. Il est donc nécessaire de pouvoir définir le cadre dans lequel évolue ce biais et des éléments adéquats permettant son estimation.

2.4 Approche mathématiques à la mesure du biais

Dans cette partie, les critères permettant de définir l'équité d'un modèle sont définis. Intuitivement, l'absence de biais se perçoit comme l'indépendance entre la variable cible et la variable sensible. En effet, l'indépendance signifierai qu'il n'y a aucun moyen que la variable sensible exerce une discrimination directe ou indirecte sur la variable cible. L'étude du biais ne devrait donc pas nécessiter de nombreux travaux. Toutefois, l'indépendance en statistique est difficile à atteindre et à définir. Et même, s'il était possible de l'atteindre, le coût sur la précision serait trop grand dans la plupart des cas. La notion d'indépendance à utiliser devient donc délicate à définir.

Une fois les critères d'indépendance présentés, les métriques permettant de mesurer le niveau d'équité seront définis. Dans la littérature, le cas de la classification avec $Y \in \{0, 1\}$ est quasi systématiquement traité. En effet, ce cas est en accord avec la majorité des cas d'usage où la mesure du biais est requise. Par exemple, dans le cas des crédits de tous types, admission dans des établissements de formations, promotion professionnelle, aide à la décision judiciaire, présence de fraude, sélection des individus suspects etc. Les cas de classifications multiclasse sont rapportés à des cas binaires en regroupant les sorties en avantageuses et désavantageuses. Des fonctions de seuils sont définies pour transformer les cas de régression en classification. L'objectif recherché étant toujours de se ramener à une distinction entre résultats avantageux et désavantageux. Le cas de la classification binaire étant le plus intuitif, les résultats sont d'abord présentés dans ce cadre. Quand cela est possible, des versions multiclasses et surtout continues sont proposées dans le but d'obtenir des résultats exploitables pour la tarification non-vie.

De plus, la variable sensible est considérée comme étant binaire. Il existe peu, voire aucune, littérature sur les autres cas de figure. Dans les cas où S à plusieurs classes, ces classes peuvent être regroupées en avantagées et désavantagées (privilégiées et non privilégiées). Dans le cas des maladies par exemple, les individus ayant des formes de maladies graves (cancer) peuvent être regroupés dans la classe "désavantagées" et les autres individus dans la classe "avantagées". Il est aussi possible de faire une étude des classes en les opposant deux à deux. Dans le cas continue, une discrétisation est envisageable. Par exemple pour la variable âge faire une discrétisation Jeune-Vieux. L'objectif étant d'étudier le biais sur une partie précise de la population, il est toujours possible de construire une variable binaire en opposant ce sous groupe de la population aux autres.

La littérature sur l'étude du biais dans les modèles est assez récente. Les références et les consensus ne sont pas explicites. Comme le décrivent des chercheurs américains en avril 2021[59] : "La croissance rapide de ce nouveau domaine a conduit à des motivations, terminologies et notations extrêmement incohérentes, présentant un sérieux défi pour le catalogage et la comparaison des définitions". Dans la 12^{eme} édition du Scientific Report de mars 2022 sur la science des données et l'intelligence artificielle, les auteurs parlent du zoo des définitions de l'équité : "le chercheur ou le praticien abordant cette facette du machine learning pour la première fois peut facilement se sentir confus et en quelque sorte perdu dans ce zoo de définitions. Ces multiples définitions saisissent différents aspects du concept d'équité mais, au summum de nos connaissances, il n'y a toujours pas de compréhension claire du paysage global où vivent ces mesures."

Ajoutez à cela le fait que les mesures soient incompatibles entre elles, que la définition du biais soit subjective et dépende fortement des objectifs ; la détection et la mitigation du biais est un domaine en expansion sur lequel il y a encore beaucoup de recherches à faire. Son importance dans le paysage économique, social et stratégique en fait toutefois un sujet de premier plan à étudier. Cette partie a pour objectif de fournir une présentation claire des notions d'équité. Elle n'est certes pas exhaustive mais sera assez complète pour donner une vision à la fois précise et globale du sujet. Elle pourra servir de référence dans la littérature francophone et permettra de disposer des outils nécessaires pour les applications actuarielles. La figure 2.8 met en avant les différentes familles d'équité qui sont étudiées dans les parties qui suivent.

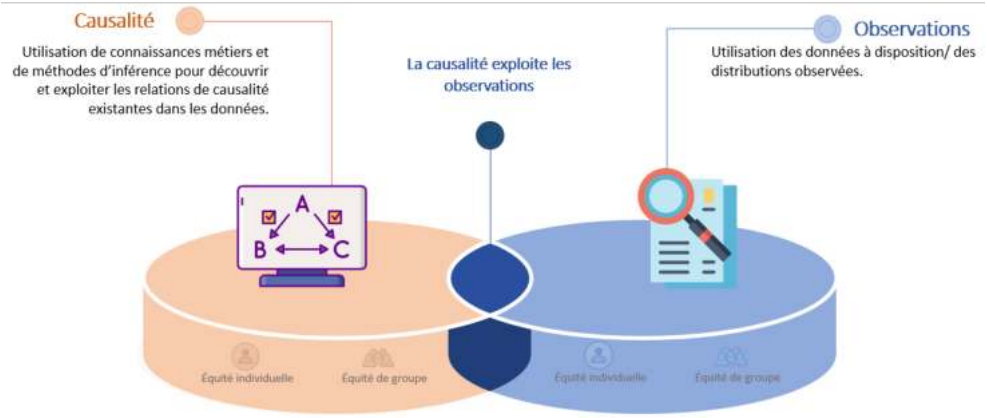


FIGURE 2.8 – Les types d'équité

Les types d'équité. L'équité se définit principalement suivant deux approches :

— Équité de groupe ou équité statistique : elle stipule que certaines statistiques doivent être les mêmes entre les individus des différents groupes formés par la variable sensible. Le but étant d'avoir le même traitement pour les individus des

groupes privilégiés et non privilégiés. C'est l'approche classique.
— Équité individuelle : elle spécifie que des individus semblables doivent être traité de la même manière. Cette approche est la plus intuitive mais aussi la plus difficile et coûteuse à mettre en place.

Au delà de ces deux distinctions, il y a une autre distinction à faire sur le cadre statistique dans lequel les critères sont définis :
— Critère basé sur les observations : ces critères ne se basent que sur les données à disposition/les distributions observées.
— Critère basé sur les causalités : ces critères sont construits en exploitant les données mais aussi les relations de causalité qui peuvent exister entre les variables. Des connaissances métiers ou des méthodes d'inférences statistiques sont utilisées dans le but de faire apparaître ces relations de causalité.

Dans chacun de ces cadres les équités individuelle et de groupe peuvent être implémentées. Le cadre causal apparaît clairement comme étant le plus intéressant. En effet, la prise en compte de la causalité peut permettre d'améliorer la modélisation statistique en certains points. Toutefois, du point de vue opérationnel, la causalité est difficile à utiliser. Il faut être capable de définir les relations de causalité entre les variables observées et prendre en compte l'effet des variables non observées. De plus, les calculs causaux se basent dans certains cas sur des évènements non observés. Ainsi, dans la littérature, le cas basé sur les observations est dominant. Il s'agit du cas par défaut. Les équités de groupe et individuelle sont souvent définies dans ce cadre en omettant complètement le cadre causal.

Dans la suite, les équités de groupe et individuelle est présentées dans le cadre des observations. Le cadre causal est présenté brièvement. Des références implicites et explicites à la causalité sont toutefois présentes tout au long des travaux suivants. La causalité étant une perspective intéressante pour comprendre les problèmes à résoudre.

2.4.1 Équité de groupe

L'équité de groupe se concentre sur le fait que pour des critères globaux données, les groupes formés par les classes de la variable sensible soient traités de la même manière. Cette définition de l'équité est la plus étudiée dans la littérature. De façon implicite, l'équité est généralement définie dans le cadre de groupe. En effet, les deux principales définitions de l'équité sont des équités de groupes. Il s'agit de la parité statistique ou démographique et des chances égalisées.

Parité démographique[21]. L'estimateur $\widehat{Y}$ respecte la condition de parité démographique par rapport à la variable sensible S si $\widehat{Y}$ est indépendante de S.

$$\mathbb{P}(\widehat{Y} \leq y|S) = \mathbb{P}(\widehat{Y} \leq y).$$

Chances égalisées[38]. L'estimateur $\widehat{Y}$ respecte la condition de chance égalisée par rapport à la variable sensible S et la cible Y si $\widehat{Y}$ est indépendante de S conditionnellement à Y.

$$\mathbb{P}(\widehat{Y} \leq y|S, Y) = \mathbb{P}(\widehat{Y} \leq y|Y).$$

Cette définition encourage l'utilisation d'informations expliquant directement Y sans passer par S. En effet, contrairement à la parité démographique, les chances égalisées permettent à $\widehat{Y}$ de dépendre de S mais uniquement à travers Y. Dans le cas binaire, ce critère requiert que les taux de faux négatifs et de faux positifs soient les mêmes à travers les classes de S.

$$\mathbb{P}(\widehat{Y} = 1|Y = 0, S = 1) = \mathbb{P}(\widehat{Y} = 1|Y = 0, S = 0)$$
$$\mathbb{P}(\widehat{Y} = 0|Y = 1, S = 1) = \mathbb{P}(\widehat{Y} = 0|Y = 1, S = 0)$$

Cette notion d'équité peut donc être relâchée en ne regardant qu'un seul pan de ces erreurs à la fois. Cela permet d'avoir les deux notions relâchées que sont l'égalité des opportunités et l'égalité prédictive.

Égalité des opportunités[38]. La vérification de ce critère nécessite que les individus, quelle que soit leur valeur de S, aient les mêmes chances de se voir attribuer la classe désavantageuse à tort. Ceci conduit à l'égalité suivante :

$$\mathbb{P}(\widehat{Y} = 0|Y = 1, S = 1) = \mathbb{P}(\widehat{Y} = 0|Y = 1, S = 0).$$

Cette mesure ne nécessite donc que l'égalité des taux des faux négatifs. Ainsi, Chances égalisées $\implies$ Égalité des opportunités.

Égalité prédictive[38]. A l'inverse de l'égalité des opportunités, ce critère requiert que les individus, quelle que soit leur valeur de S, aient les mêmes chances de se voir attribuer la classe avantageuse à tort. D'où l'égalité suivante :

$$\mathbb{P}(\widehat{Y} = 1|Y = 0, S = 1) = \mathbb{P}(\widehat{Y} = 1|Y = 0, S = 0).$$

Cette mesure ne nécessite donc que l'égalité des taux de faux positifs. Ainsi, Chances égalisées $\implies$ Égalité prédictive. Le choix entre ces deux versions dépend du problème traité et l'aspect sur lequel doit être le plus porté l'équité.

En plus de ces deux définitions majeures, deux autres définitions permettent de compléter le spectre des types de définitions de l'équité. Il s'agit de la parité prédictive et du traitement disparate.

Parité prédictive[16]. La parité prédictive est vérifiée quand par rapport aux estimations faites ou décisions prises $\widehat{Y}$, Y est indépendante de S.

$$\mathbb{P}(Y \leq y|S, \widehat{Y}) = \mathbb{P}(Y \leq y|\widehat{Y}).$$

Traitement disparate[83]**.** Un modèle souffre de traitement disparate si sachant X, $\widehat{Y}$ et S sont dépendants. En d'autres termes, si les prédictions sont différentes pour des individus ayant les mêmes caractéristiques X avec des valeurs de S différentes.

$$\widehat{Y} \mid X \perp\!\!\!\perp S$$

Pour mieux comprendre ces critères, il est important de comprendre le type d'équité qu'ils reflètent. Ces critères peuvent être regroupés en 3 sous groupes : les critères d'indépendance, de suffisance et de séparation.

L'indépendance

Elle stipule que les résultats $\widehat{Y}$ doivent être indépendants de S. La parité démographique est un cas d'indépendance. Ainsi, les chances d'obtenir des valeurs de $\widehat{Y}$ doivent être les mêmes quel que soit S. La figure 2.9, inspirée des travaux de Castelnovo et al.[12], permet de voir à quoi cela correspond dans un exemple de sélection de risque. La variable sensible est le genre et l'objectif est d'accepter les bons risques et de rejeter les risques aggravés (ou étudier la possibilité d'appliquer une surprime).

FIGURE 2.9 – Illustration d'un modèle respectant la parité statistique

Les critères basés sur l'indépendance n'utilisent pas les Y. Ainsi, seules les variables X, $\widehat{Y}$ et S sont exploitées. Dans le cas de la tarification par exemple cela se traduit par le fait que pour tous les âges, la prime soit identique en prenant l'âge comme variable sensible. Ainsi, l'égalité est imposée entre les âges sans tenir compte de quelconque autres informations. Il apparaît donc que pour ces critères, il est nécessaire de traiter les groupes de manières différentes dans le but de les mettre sur le même pieds. En effet, si le risque porté par un individu d'un certain âge est plus grand, il faudra être moins regardant à ce niveau pour permettre aux individus d'avoir les mêmes primes quel que soit leur âge. Cette approche est contre-intuitive avec l'idée d'équité qui est celle de traiter les individus sans tenir compte de la variable sensible. Il a été discuté dans la littérature que ce comportement des critères d'indépendance pouvait, dans certains cas, conduire à une amplification du biais. En effet, "les discriminations positives" appliquées dans le but d'atteindre l'équité peuvent creuser plus grandement le fossé entre les classes de S.

Cette notion est donc à appliquer en priorité dans les cas où le biais historique est significativement présent dans les données. En effet, la variable Y n'est pas utilisable car biaisé. Même si une relation semble exister entre S et Y, elle est peut être le fruit du biais. Il peut aussi arriver que toute relation entre S et Y soit tout simplement jugée inadmissible ou que des variables explicatives indépendantes de S en théorie soient suffisantes pour expliquer le phénomène observé.

La notion d'indépendance se présente aussi sous une forme conditionnelle. Ainsi, en lieu et place de l'indépendance totale entre $\widehat{Y}$ et S, il s'agit d'une indépendance conditionnée par une autre variable explicative :

$$\widehat{Y} \perp\!\!\!\perp S|X_1.$$

Cette approche permet de faire disparaître les relations avec S par rapport à une variable donnée. Ainsi, si X_1 est une variable représentant le niveau de risque, $\widehat{Y}$ serait indépendante de S pour chaque niveau de risque donné. Toutefois, il faut s'assurer que la variable X_1 ne soit pas biaisée. En effet, si cette variable est biaisée, les résultats le seront aussi. Le choix de la variable peut aussi introduire du biais en favorisant des pans différents de la population.

Les 2 prochaines notions sont la séparation et la suffisance. Ces notions intègrent les informations contenues dans Y pour mesurer l'équité d'un modèle.

Séparation

Elle est une indépendance conditionnelle par rapport à la vraie valeur de la variable cible Y. Ainsi, la séparation s'écrit :

$$\widehat{Y} \perp\!\!\!\perp S|Y.$$

Cela signifie que toute différence de traitement entre les classes de S doit être justifiée par la valeur de Y. Ainsi, en reprenant le cas de la prime et de l'âge, la prime ne sera pas la même à tous les âges comme dans le cas de l'indépendance. Dans le cas de la séparation, la prime peut être différente pour chaque âge si des facteurs de risques indépendants de l'âge le justifient. Les chances égalisées sont une forme de séparation. La figure 2.10 illustre cela dans le cas discret de la sélection du risque.

FIGURE 2.10 – illustration d'un modèle respectant les chances égalisées

Cette approche semble donc satisfaisante car elle permet d'éliminer l'effet de la variable sensible tout en préservant les informations contenues dans Y. Toutefois, comme discuté dans la partie conditionnement, cette approche permet de réduire le biais seulement dans le cas où la variable Y ne contient pas de biais. En effet, si cette variable est biaisée, le biais sera transmis à $\widehat{Y}$.

Cette notion doit donc s'appliquer en priorité dans les cas où le biais historique est absent ou très faible. Il faut que Y soit une variable de confiance permettant de réduire le biais.

Suffisance

La séparation intègre les informations par rapport à Y en se plaçant du point de vue de la vraie valeur Y. La suffisance quant à elle, regarde l'équité par rapport aux décisions du modèle. Elle impose la parité entre les individus ayant les mêmes valeurs prédites. L'évaluation de Y sachant la valeur de $\widehat{Y}$ attribuée ne doit pas dépendre de S. Ainsi, la suffisance revient à :

$$Y \perp\!\!\!\perp S|\widehat{Y}.$$

Par exemple dans le cas de la sélection du risque, cela revient à dire que sachant la décision d'accepter ou non le risque, le genre n'a aucun effet sur l'évaluation des évènements risque réellement aggravé ou bon risque. La parité prédictive est une forme de suffisance. L'intérêt d'une telle approche réside dans le fait que pour des individus inconnus, la vraie valeur de Y n'est pas encore observée. De plus, dans les processus incluant des sélections en entrée, la variable Y est toujours sujette à un certain biais de sélection. En effet, seuls les individus préalablement choisis ont pu être observés. Faire le raisonnement en partant de $\widehat{Y}$ permet de réduire ce biais.

La définition de ces différentes notions d'indépendance s'explique par le fait que l'indépendance stricte n'est soit pas atteignable soit pas envisageable/souhaitable. Il est donc indispensable d'avoir un champ de définition plus vaste permettant de mesurer des facettes différentes du biais. L'équité individuelle apporte une autre nuance, elle prône une vision locale du biais contrairement à la vision globale proposée par l'équité de groupe.

2.4.2 Équité individuelle

L'équité individuelle est aussi appelée équité basée sur la similarité. Son principe général est que des individus similaires doivent avoir des prédictions similaires. Ainsi, contrairement à l'équité de groupe, aucun agrégat n'est calculé. Les individus sont comparés les uns aux autres.

Une des idées qui vient intuitivement est l'équité par omission. La variable cible n'apparaît plus explicitement dans le modèle, et donc les individus identiques ayant des valeurs différentes de la valeur sensible ont exactement les mêmes prédictions. Comme expliqué plus haut, cette méthode ne tient pas compte des relations qui existent entre X et S. Deux individus similaires dans (X, S) (indépendamment de S) ne peuvent pas tout simplement se définir comme étant des individus ayant des valeurs identiques de X. Dans le cas du genre par exemple, le changement de genre induit directement un changement

dans la distribution des valeurs prises par d'autres variables explicatives. Cela est aussi vrai pour la maladie, le handicap, l'âge etc.
Il apparaît donc que la définition de la notion de similarité entre les individus sur X est difficile. Dans le cas où une définition puisse être construite, il faudra faire attention à ce que cette définition ne soit pas source de biais. Ces difficultés font de l'équité individuelle un pan de l'équité très peu développé dans la littérature.

Des approches ont toutefois été proposées. Dwork et al[21], formulent pour la première fois l'équité individuelle sous la forme d'une condition Lipschitzienne :

$$d_Y(\widehat{y}_i, \widehat{y}_j) < \lambda d_X(x_i, x_j),$$

avec d_Y et d_X respectivement des distances sur l'espace de la variable cible et sur celui des variables explicatives, et λ une constante. La distance d_Y peut se définir aisément comme étant $d_Y = |\widehat{Y}_i - \widehat{Y}_j|$. La distance d_X est plus complexe à définir. Elle doit être à la fois cohérente avec les données et équitable. Une des distances proposée par Dwork et al. est :

$$d_X = 1 - \frac{1}{n}\left(\sum_{i=1}^{n}\left|\widehat{y}_i - \frac{1}{k}\sum_{x_j \in \mathbb{V}_{\mathrm{KNN}}(x_i)} \widehat{y}_j\right|\right).$$

Cette mesure revient à utiliser le voisinage défini par la méthode des k plus proches voisins ($\mathbb{V}_{\mathrm{KNN}}$) comme étant la mesure de similarité. Cette approche a toutefois l'inconvénient de nécessiter la définition d'une distance dans la méthode de k plus proches voisins. Le problème du choix de la distance est donc reporté mais pas résolu. De plus, cette méthode ne permet pas de mitiger le biais dans tous les cas. En effet, dans le cas des maladies graves, si les voisins d'un malade sont d'autres malades, les résultats seront jugés équitables alors que le biais entre malades et non malades peut exister sans être détectable dans le voisinage immédiat de l'individu. Ces approches basées sur les distances sont appelées équité par conscience. Dans des travaux plus récents[40, 43], des chercheurs se sont posés la question de savoir si la mesure de similarité ne devait pas être construite au cas par cas. En effet, cela permettrait de capter toutes les particularités du problème traité en mettant l'accent sur les variables les plus importantes et en décidant ce qui est admissible comme discrimination et ce qui ne l'est pas.

Une autre approche pour pallier les limites de l'équité par omission serait d'avoir recours à la suppression ou au nettoyage. La suppression consiste à retirer des données toutes variables présentant un lien significatif avec S. Ainsi, un nouveau $\tilde{X}$ est construit de telle sorte qu'aucune relation n'existe entre lui et S. Dans la pratique, cela conduirait à une dégradation significative de la qualité des modèles du fait de la grande perte d'informations. Le principe de nettoyage a donc été introduit dans le but d'obtenir un $\tilde{X}$ ne contenant aucune information sur S en conservant les informations utiles à la prédiction. Cela peut se faire par exemple en construisant un modèle permettant de reconstruire X tout en le pénalisant pour le rendre indépendant à S[56].

L'équité individuelle offre une bonne transition vers les équités basées sur la causalité. En effet, la causalité offre un cadre adéquat en théorie pour mesurer les effets des variables les unes sur les autres.

2.4.3 Équité basée sur la causalité

L'introduction de la causalité dans les analyses statistiques a des avantages non négligeables. Elle permet d'introduire les connaissances métiers et de faire face aux limites des analyses basées sur les corrélations. Elle permet aussi d'avoir un cadre théorique dans lequel il est possible de répondre à la question de l'équité : quelle aurait été la valeur de $\widehat{Y}$ si S avait une valeur différente ? Quelques travaux ont essayé, sur des cas d'usage précis, de définir et d'appliquer l'équité basée sur la causalité[53, 34, 82]. Ce cadre permet aussi de prendre en compte l'effet de variable sensible non observée. Cela peut être le cas des origines par exemple. En règle générale, avant de faire des travaux dans le cadre causal, il est nécessaire de construire un graphe causal. Ce graphe met en avant les relations entre toutes les variables en jeu, observées et parfois non observées. Ce graphe est construit à l'aide de la connaissance du problème traitée avec l'appui de méthodes d'inférence statistique. La figure 2.11 fournit un exemple de graphe causal sur la base de données Adult UCI dans laquelle le revenu est la variable cible.

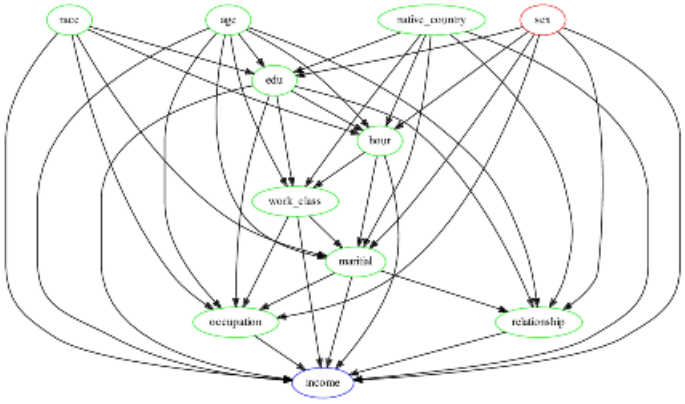


FIGURE 2.11 – Exemple de graphe causal ; Auteur : Lu Zhang

Bien que très prometteuse, l'équité causale est encore très peu développée. Cela s'explique par le fait que la causalité soit en elle-même un domaine des statistiques en croissance. De plus, elle se base sur des hypothèses fortes de relations entre les variables. Ces hypothèses nuisent à la robustesse de cette approche. Un autre point est que la construction du graphe causal peut être très difficile. Dans le cas où il serait construit, il existe une infinité de modèles causaux pouvant être construits à partir du graphe.
La causalité souffre des inconvénients de ces avantages. Étant une approche plus complète, elle fait souvent référence à des distributions, des évènements ou des variables non observés et/ou difficilement estimables.

Les théories causales et contre factuelles proviennent des travaux de Judea Pearl. Il a introduit le "do-calculus" [1] qui est utilisé pour définir les critères d'équité[64, 63]. Les critères causaux suivants sont définis :

1. do : faire, calculus : calcul ; peut se traduire comme les calculs de l'action

Équité contre-factuelle
Un modèle respecte ce critère si :

$$\mathbb{P}(\widehat{Y}_{S\leftarrow 1} = 1|X = x, S = 1) = \mathbb{P}(\widehat{Y}_{S\leftarrow 0} = 1|X = x, S = 1).$$

Cela signifie qu'en prenant un individu avec $X = x$ et $S = 1$, et le même individu avec $S = 0$, ils doivent avoir les mêmes chances d'obtenir la valeur $\widehat{Y} = 1$. Ce critère d'équité est le critère le plus individuel.
Il apparaît clairement que l'individu avec $X = x$ et $S = 0$ peut ne pas être observé. Ainsi, en plus du graphe causal, un modèle structurel causal doit être introduit (Structural Causal Model, SCM). Ce modèle doit être construit de manière à pouvoir décrire les effets des variables les unes sur les autres.

Équité contre-factuelle espérée
Un modèle respecte ce critère si :

$$\mathbb{P}(\widehat{Y}_{S\leftarrow 1} = 1|S = 1) = \mathbb{P}(\widehat{Y}_{S\leftarrow 0} = 1|S = 1).$$

Ainsi, dans la population, les individus avec $S = 1$ et les individus contraints de passer de $S = 1$ à $S = 0$ doivent avoir en moyenne les mêmes chances la valeur $\widehat{Y} = 1$. Ce critère d'équité est le moins individuel.

Entre ces deux critères se trouvent les critères conditionnés par rapport à d'autres variables des données. De plus, de par sa structure plus riche, l'équité causale permet de spécifier les relations qui seront admises entre X, S et $\widehat{Y}$ et celles qui ne le seront pas.

La définition de ces critères se présente de la même manière que les définitions de la partie sur les observations. Les différences résident dans les méthodes utilisées dans le but d'estimer les probabilités et les espérances. Dans l'univers causal, en plus de l'approche contre-factuelle, une approche basée sur les interventions est envisageable. La dernière définition deviendrait par exemple :

$$\mathbb{P}(\widehat{Y} = 1|do(S = 1)) = \mathbb{P}(\widehat{Y} = 1|do(S = 0)).$$

Cette définition peut se traduire comme étant d'imposer l'égalité des chances d'obtention de la valeur $\widehat{Y} = 1$ pour un individu tiré aléatoirement de la population et contraint d'avoir la valeur $S = 1$ et un autre individu tiré aléatoirement et contraint d'avoir la valeur $S = 0$. Ces deux nuances de la causalité sont assez subtiles à distinguer.
Dans la suite de ces travaux, le "do-calculus" et les graphes causaux ne sont pas utilisés puisque ces éléments trop spécifiques nuiraient à la capacité de généralisation des travaux effectués. La causalité est envisagée d'un point de vue moins stricte : construire des métriques prenant en compte les effets des variables les unes sur les autres sans construire de structures ou de graphes causaux.
La figure 2.12 résume la décomposition des critères de définitions de l'équité.

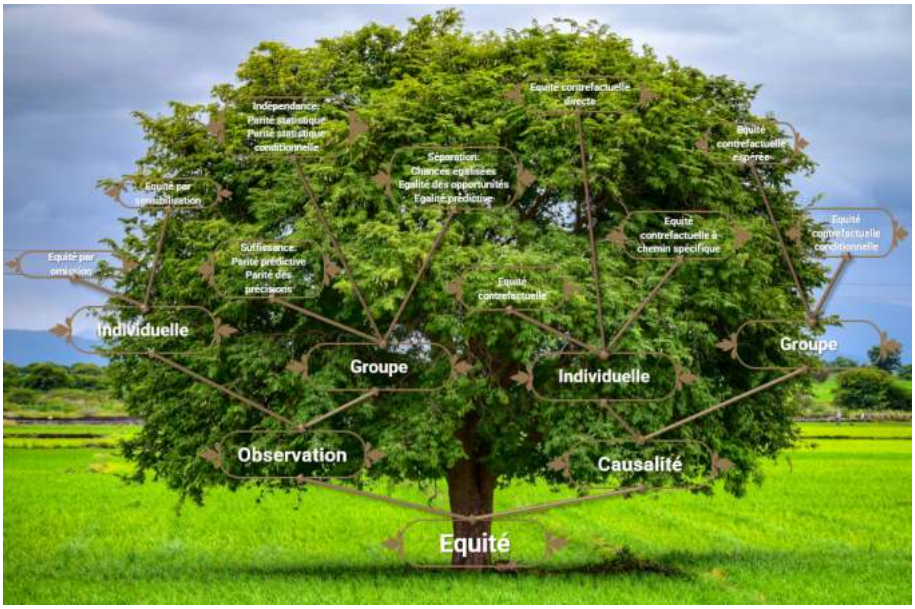


FIGURE 2.12 – L'arborescence des critères de définitions de l'équité.

2.4.4 Équité de groupe contre équité individuelle : Analyse du spectre d'équité basée sur les observations

L'équité individuelle stipule que les individus similaires doivent avoir des résultats similaires. Ainsi, par définition, elle permet d'éviter le traitement disparate. Le traitement disparate est un aspect important de l'équité. En effet, il est assez intuitif d'attribuer les mêmes résultats aux individus semblables sur X sans faire de distinction par classe de S.

Toutefois, la parité de groupe ne contraint pas à l'obtention de cette condition d'équité. En effet, l'équité de groupe stipule qu'en moyenne les individus de groupes différents doivent être traités de la même manière suivant certaines mesures. Il peut donc arriver que des discriminations restent présentes à l'intérieur des groupes. Des individus similaires se retrouvent donc avec des résultats différents. Par exemple, une certaine "discrimination positive" est observable dans la parité statistique. Les individus des différents groupes sont traités différemment dans le but de permettre de créer une équité en moyenne. Comme discuté dans la section équité de groupe, cette approche est contre-intuitive et peut conduire à une exacerbation du biais. Elle est en opposition à la définition de l'équité individuelle.

Deux facteurs externes sont importants à prendre en compte. Ce sont la perte d'information et la capacité d'implémentation liée aux méthodes d'équité. En effet, la construction d'une notion d'équité parfaite mais très peu performante et difficile à implémenter n'aura aucun intérêt pratique. Par exemple, pour rendre les notions de groupes plus sensibles au traitement disparate, il est possible d'utiliser la "suppression" et de ne prendre en compte aucune variable de X liée à S. Ainsi, aucune discrimination de groupe ou indivi-

duelle ne pourra être présente dans les résultats. Cependant, cette approche conduirait sans doute à des performances significativement plus faibles sauf dans le cas dégénéré où $X \perp\!\!\!\perp S$. Il semble donc y avoir un échange entre la quantité d'information conservée sur S (et donc l'équité) et la capacité à imposer un traitement individuel non disparate. Plus la quantité d'information conservée sur S est grande, plus il est simple d'imposer un traitement équitable sur des individus semblables de valeurs de S différentes. Plus le modèle est équitable par rapport à S, (moins d'informations), plus il est difficile de s'assurer que les individus soient tous traités équitablement par rapport à S. Entre ces deux extrêmes se trouve les équités de groupes conditionnelles. Ces approches n'imposent pas d'indépendance stricte et donc utilisent plus d'informations sur S. Ils sont donc meilleurs dans la prise en compte du traitement disparate.

La figure 2.13 inspirée des travaux de Castelnovo et al.[12] présentent les critères d'équité avec en abscisse le niveau d'équité individuelle du modèle et en ordonnée la quantité d'information de S utilisée.

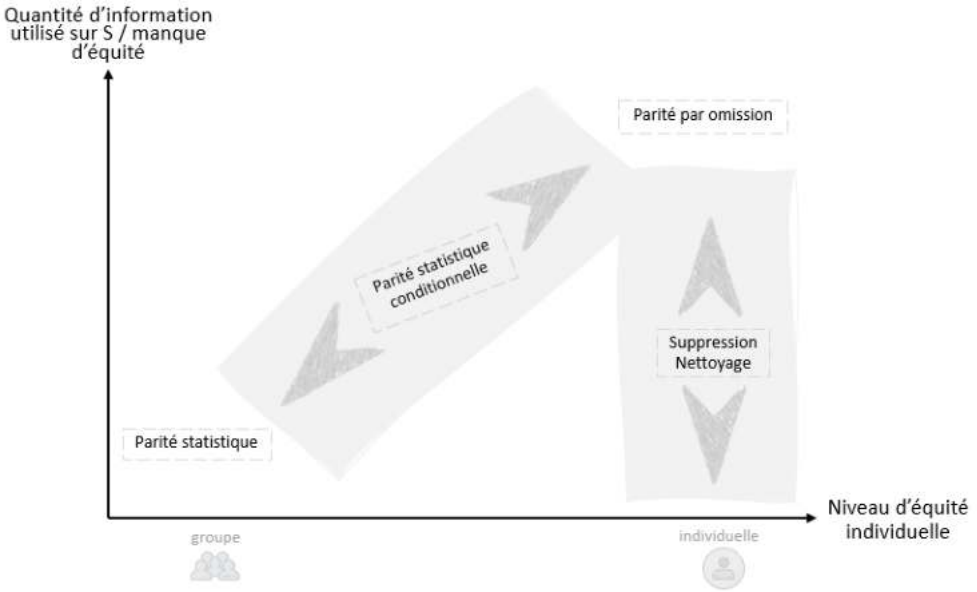


FIGURE 2.13 – Positionnement des critères d'équité dans le plan individualisme du critère et quantité d'informations de S utilisée.

Il est intéressant de remarquer qu'en utilisant des méthodes de nettoyage, l'équité individuelle est mise à mal. En effet, ces méthodes cherchent à reconstruire X en conservant toutes les informations utiles et en imposant une indépendance avec S. En faisant ceci, la quantité d'information utilisée sur S baisse. En contrepartie, des individus ayant des X semblables mais des S différents seront, dans certains cas, considérés comme des individus différents.

En regardant les notions d'équité à travers le prisme de la performance, une relation prévisible apparaît : la performance est positivement corrélée avec la quantité d'information de S utilisée (l'absence d'équité). Un des défis de la mitigation du biais réside dans

le dilemme performance-équité. Ainsi, la méthode la moins équitable, qui ne tient pas compte de l'existence d'interdépendance entre X et S, devrait être la plus performante (équité par omission). La méthode imposant une forte indépendance avec S est à priori la moins performante (parité statistique). Entre ces deux extrêmes se trouvent les méthodes de parité conditionnelle, de suppression et de nettoyage. Ce sont des méthodes qui conservent une plus grande partie de l'information contenue dans S et donc sont plus performantes. Cela donne la représentation visible sur la figure 2.14.

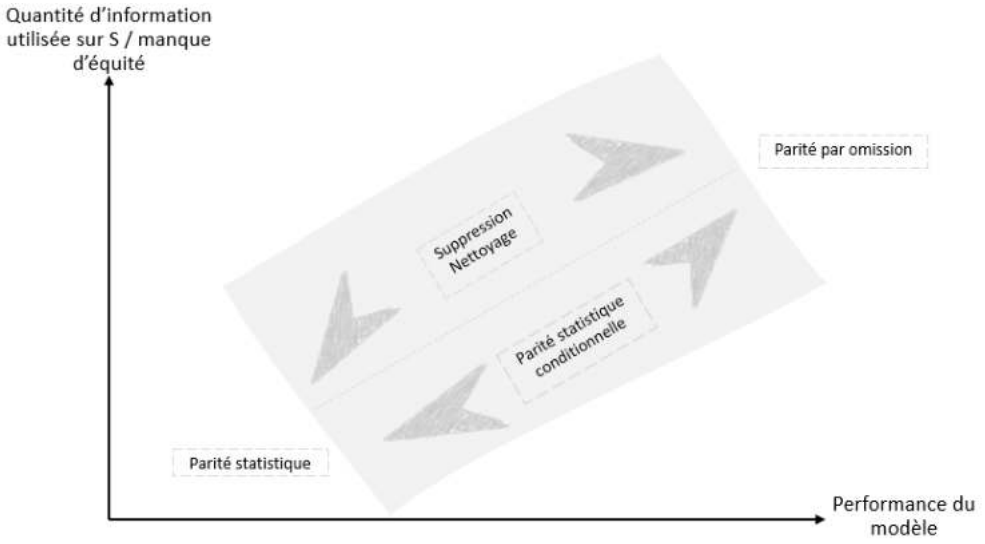


FIGURE 2.14 – Positionnement des critères d'équité dans le plan performance du modèle et quantité d'informations de S utilisée.

Certaines sources[12, 6] s'accordent à dire que l'équité de groupe et l'équité individuelle ne sont pas en conflit mais sont plutôt les deux extrêmes d'une droite imaginaire où se trouvent les notions d'équité. Un des principaux leviers est la quantité d'informations sur S qui est conservée dans les données. Cette question devrait vraisemblablement être abordée du point de vue du problème à résoudre. La question deviendrait, du point de vue du cas d'usage : quel est le niveau acceptable d'interdépendance avec S ? Quelles sont les variables acceptables ? Quelles sont les contraintes commerciales, légales ou éthiques ?

Ainsi, dans le but de faire la mitigation du biais, il faut :

1. Choisir le cadre d'application de la mesure du biais : cadre d'observation ou causal ;
2. Choisir la notion d'équité à appliquer : équité individuelle ou de groupe ;
3. Choisir les métriques adéquates à la mesure suivant le cas d'usage.

La prochaine étape est donc de définir les métriques exactes qui permettent de mesurer le biais.

2.5 Les métriques de mesures de biais

Pour alléger les définitions, les notations suivantes sont introduites :
— VP (Vrai Positif) : $\widehat{Y} = 1 \mid Y = 1$
— VN (Vrai Négatif) : $\widehat{Y} = 0 \mid Y = 0$
— FP (Faux Positif) : $\widehat{Y} = 1 \mid Y = 0$
— FN (Faux Négatif) : $\widehat{Y} = 0 \mid Y = 1$

$\widehat{Y}/Y$	0	1
0	VN	FN
1	FP	VP

TABLE 2.5 – Matrice de confusion : classification binaire

Les métriques classiques de classification s'obtiennent directement en exploitant la matrice de confusion :
— Le rappel : $\frac{VP}{VP+FN}$, permet de mesurer la capacité du modèle à distinguer les positifs. Il est noté TVP (Taux de Vrai Positif) pour désigner sa formulation générale : $\mathbb{P}(\widehat{Y} = 1 \mid Y = 1)$
— La précision : $\frac{VP}{VP+FP}$, permet de mesurer la part d'individus réellement positifs parmi ceux classés positifs par le modèle. Elle est notée VPP (Valeur Prédictive Positive) pour désigner sa formulation générale : $\mathbb{P}(\widehat{Y} = 1 \mid Y = 1)$.
— La spécificité : $\frac{VN}{VN+FN}$, permet de mesurer la part d'individus réellement négatifs parmi ceux classés négatifs par le modèle. La notation TFP (Taux de Faux Positifs) est utilisée. Elle désigne la part de négatifs ayant été classée positifs par le modèle. TFP = 1 - spécificité = $\mathbb{P}(\widehat{Y} = 1|Y = 0)$.
— Le taux de succès : $\frac{VP+VN}{VP+VN+FP+FN}$, permet de mesurer la probabilité que le modèle attribue la bonne classe à un individu quelconque.
— F1-score : $\frac{2 \times \text{ rappel } \times \text{ précision}}{\text{précision } + \text{ rappel}}$.

Le taux de succès est la métrique usuelle. Elle a montré ses limites pour des problèmes de classifications où les classes n'avaient pas le même poids. Les autres métriques, plus spécifiques, ont été introduites pour mesurer les performances du modèle sur des sous parties des données.

Pour la mitigation du biais, au-delà de la relation entre $\widehat{Y}$ et Y, il est indispensable de prendre en compte les relations réciproques avec S. Les métriques habituelles ne sont donc plus suffisantes car il faut être en capacité de prendre en compte les distributions conditionnellement à S. Il faut donc construire des matrices de confusion en fonction de la valeur de S. Ainsi, les notations suivantes sont introduites pour $s \in \{0, 1\}$:
— TVP_s : $\mathbb{P}(\widehat{Y} = 1|Y = 1, S = s)$, taux de vrais positifs sachant S.
— TFP_s : $\mathbb{P}(\widehat{Y} = 1|Y = 0, S = s)$, taux de faux positifs sachant S.
— VPP_s : $\mathbb{P}(Y = 1|\widehat{Y} = 1, S = s)$, valeur prédictive positive sachant S.

Dans ce qui suit, la classification avec $Y \in \{0, 1\}$ est présentée en premier. Une adaptation

au cas continue et/ou multiclasse est introduite si possible. Ω_Y dénote l'ensemble des valeurs prises par Y. Les métriques prenant en compte la variable sensible S peuvent être distinguées en deux groupes majeurs. Les métriques construites après la modélisation et les métriques construites avant la modélisation.

2.5.1 Les métriques construites après la modélisation

Les métriques construites après la modélisation sont des métriques qui exploitent la relation entre $\widehat{Y}$ et S en utilisant par moment la variable Y.

La p-rule.

$$\text{La p-rule} : \min\left(\frac{\mathbb{P}(\widehat{Y}=1|S=1)}{\mathbb{P}(\widehat{Y}=1|S=0)}, \frac{\mathbb{P}(\widehat{Y}=1|S=0)}{\mathbb{P}(\widehat{Y}=1|S=1)}\right).$$

Ainsi, si p-rule = 1, le modèle respecte la parité statistique car la probabilité d'avoir $\widehat{Y}=1$ ne dépend pas de la valeur de S. Une différence peut aussi être calculée à la place du ratio :

$$\mathbb{P}(\widehat{Y}=1|S=0) - \mathbb{P}(\widehat{Y}=1|S=1)$$

Impact disparate.

$$\text{ID} : |\mathbb{P}(\widehat{Y}=1|S=1) - \mathbb{P}(\widehat{Y}=1|S=0).$$

Plus la différence est petite, plus le modèle est considéré comme juste suivant le critère de parité statistique. Une version sous forme de ratio est aussi introduite :

$$\text{ID} : \frac{\mathbb{P}(\widehat{Y}=1|S=1)}{\mathbb{P}(\widehat{Y}=1|S=0)}.$$

Si la valeur de ce ratio est comprise entre 80% et 120%, le modèle peut être admis comme étant acceptable. 100% étant la valeur idéale.

+ *Cas multiclasse :*

$$\forall y \in \Omega_Y,\ \frac{\mathbb{P}(\widehat{Y}=y|S=1)}{\mathbb{P}(\widehat{Y}=y|S=0)} \in [0,80; 1,20].$$

+ *Cas continu :* ratio des moyennes de Y pour chaque classe de S :

$$\frac{\mathbb{E}(\widehat{Y}|S=1)}{\mathbb{E}(\widehat{Y}|S=0)} \in [0,80; 1,20].$$

Maltraitance disparate. L'égalité des chances peut se mesurer en observant les différences entre les taux de faux positifs et de faux négatifs pour chaque modalités de la variable sensible :

$$M_{TFP} : |\mathbb{P}(\widehat{Y} = 1|Y = 0, S = 1) - \mathbb{P}(\widehat{Y} = 1|Y = 0, S = 0)|$$

$$M_{TFN} : |\mathbb{P}(\widehat{Y} = 0|Y = 1, S = 1) - \mathbb{P}(\widehat{Y} = 0|Y = 1, S = 0)|$$

Plus les valeurs sont proches de 0, plus le modèle est équitable suivant le critère des chances égalisées. Ce critère se réécrit souvent sous une forme appelée Moyenne des Différences de Chances :

$$MDC = \frac{(TFP_0 - TFP_1) + (TVP_0 - TVP_1)}{2}.$$

Plus la métrique est proche de 0, plus le modèle est juste suivant le critère d'égalité des chances. Ainsi, le modèle est considéré comme équitable si les chances d'être bien ou mal classé sont les mêmes quelle que soit la valeur de S.

Les autres métriques de différences entre les classes. Au lieu de regarder les taux de faux positifs et de faux négatifs simultanément, il est possible de regarder les différences de précisions entre les deux classes, les différences de taux d'erreur, le ratio de taux de faux négatifs sur taux de faux positifs. L'objectif étant de quantifier toutes différences de traitements entre les classes de la variable sensible.

+ *Cas multiclasse :* les calculs peuvent être fait suivant chacune des classes de Y prise individuellement puis comparés aux autres. Par exemple, les taux d'erreurs peuvent être calculés pour chacune des valeurs de Y et comparés les unes aux autres. Les taux de faux positifs peuvent être calculés en se ramenant dans le cas binaire. Par exemple, si Y prend 3 valeurs $\{0, 1, 2\}$, le taux de faux positifs en 0 peut se calculer en fixant $Y = 0$ quand $Y = 0$ et $Y = 1$ quand $Y \in \{1, 2\}$.

+ *Cas continu :* une adaptation au cas continu n'est pas possible pour ces métriques. Le seul calcul envisageable est celui de la comparaison des ratios de valeur moyenne prédite sur valeur moyenne observée dans les données.

$$\frac{\mathbb{E}(\widehat{Y}|S = 1)}{\mathbb{E}(Y|S = 1)}.$$

Dans la continuité de ce ratio comparant prédictions et observations, des métriques peuvent être construites pour s'assurer que le modèle n'introduit pas de biais dans les résultats. Cela peut se faire par exemple en calculant :

$$\frac{\#\{\widehat{Y} = 1|S = 1\}}{\#\{Y = 1|S = 1\}} - \frac{\#\{\widehat{Y} = 1|S = 0\}}{\#\{Y = 1|S = 0\}}.$$

En partant de l'hypothèse que les données sont non biaisées, ces métriques permettent de s'assurer que les sorties $\widehat{Y}$ ne le sont pas aussi. Ils permettent également de s'assurer que le modèle n'est pas biaisé de manière volontaire en s'assurant que le comportement du modèle est cohérent avec celui observé dans les données.

Différence contre-factuelle. Cette métrique est une version simplifiée de l'étude de l'équité contre-factuelle. Elle est basée sur le principe que si le modèle est équitable par rapport à S, modifier la valeur S ne devrait pas changer les résultats obtenus. C'est une version simplifiée car elle ne fait pas appel à un graphe causal ou au "do-calculus". Dans cette approche, pour chaque individu de la classe $S = 1$, il faut remplacer la valeur de S par $S = 0$ et ensuite refaire toutes les prédictions. Le nombre d'individu ayant vu leur valeur de $\widehat{Y}$ modifiée est noté $N^{S=1}_{do(S=0)}$. $N^{S=0}_{do(S=1)}$ est définie d'une manière similaire.

Avec ces notations, Galhotra et al.[29] introduisent le score de discrimination causal :

$$SDC = \frac{N^{S=1}_{do(S=0)} + N^{S=0}_{do(S=1)}}{N},$$

où N est le nombre d'individus dans les deux classes. Il est aussi possible de mesurer la robustesse des classes de S. En notant

$$r_{S=1} = \frac{N^{S=1}_{do(S=0)}}{\#\{S=0\}} \text{ et } r_{S=0} = \frac{N^{S=0}_{do(S=1)}}{\#\{S=1\}},$$

la sensibilité contre-factuelle peut être définie[18] :

$$SC_1 = r_{S=0} - r_{S=1} \text{ et } SC_2 = \frac{r_{S=0} - r_{S=1}}{r_{S=0} + r_{S=1}}.$$

Si les quantités SC_1 et SC_2 sont strictement positives, alors la classe $S = 0$ est moins robuste que la classe $S = 1$. Cette différence de robustesse est une forme de biais.

\+ *Cas multiclasse :* comme pour les métriques précédentes, il est possible de tenter une adaptation en comparant les classes deux à deux ou en les regroupant.

\+ *Cas continu :* la seule possibilité dans ce cas est de comparer les distributions de $\widehat{Y}$ après la modification des valeurs de S. Ces distributions pourraient être comparées avec les distributions avant la modification.

Flip-test. Cette métrique permet de mesurer l'équité causale individuelle. Il prend mieux en compte les contraintes du cadre causal. Basée sur les travaux de Dwork et al.[21] et introduite par Black et al.[7], cette métrique utilise des modèles GAN [2] dans le but de concevoir, en partant d'un individu avec une valeur fixée de S, un individu

2. Generative Adversarial Network. Ce sont des réseaux neuronaux spécialisés dans la génération sous contraintes d'individus.

similaire avec un S différent. Cette approche permet de prendre en compte le fait qu'en changeant la valeur de S, les valeurs de X peuvent aussi être modifiées. En partant de ces travaux, Das et al.[18] proposent une approche simplifiée dans laquelle au lieu de générer des individus avec des modèles GAN, la méthode des k plus proches voisins est utilisée pour détecter les individus de la classe opposée les plus proches. Le test s'articule de la manière suivante :

- Choisir la classe par rapport à laquelle le biais sera mesuré. Pour illustration, la classe $S = 1$ est utilisée et est considérée comme la classe désavantagée. De plus, la classe $Y = 0$ est considérée comme le résultat désavantageux. Par exemple, être rejeté par l'assureur.
- Pour chaque individu de la classe désavantagée $S = 1$, ayant une prédiction défavorable $Y = 0$, il faut trouver ses plus proches voisins de la classe avantagée $S = 0$ et vérifier quelle est la prédiction dominante parmi ces voisins. Le nombre d'individu ayant un voisinage de classe avantagée ayant une prédominance de prédiction favorable est noté F^+.
- Pour chaque individu de la classe désavantagée $S = 1$, ayant une prédiction favorable $Y = 1$, il faut trouver ses plus proches voisins de la classe avantagée $S = 0$ et vérifier quelle est la prédiction dominante parmi ces voisins. Le nombre d'individu ayant un voisinage de classe avantagée ayant une prédominance de prédiction défavorable est noté F^-.
- Ainsi,

$$FT = \frac{F^+ - F^-}{\#\{S = 1\}}.$$

Plus la métrique FT est proche de 0, moins le modèle est discriminant. Plus elle est proche de 1 plus le modèle est biaisé en faveur de $S = 0$ et plus elle est proche de -1 plus le modèle est biaisé en faveur de $S = 1$.

+ Cas multiclasse : regrouper les modalités, les comparer deux à deux ou un contre tous.

+ Cas continu : en partant de la forme générale proposée par Dwork et al. nécessitant le calcul de la moyenne des différences de prédiction entre l'individu et son voisinage (voir section équité individuelle), une adaptation serait de ne considérer que les individus de la classe désavantagée et leur voisinage de la classe avantagée. Dans le cas du genre par exemple le calcul prendrait la forme suivante :

$$\tilde{FT} = \frac{1}{N_F}\left(\sum_{i=1}^{N_F}\left[\widehat{y}_i - \frac{1}{k}\sum_{x_j \in \mathbb{V}^M_{\mathrm{KNN}}(x_i)} \widehat{y}_j\right]\right),$$

où N_F est le nombre de femmes et $\mathbb{V}^M_{\mathrm{KNN}}(x_i)$ est le voisinage masculin de la femme i.

Les mesures introduites plus haut sont des métriques calculables une fois le modèle construit. Toutefois, avec la connaissance de Y et de S, il est possible de mesurer le biais existant à priori dans les données.

2.5.2 Les métriques construites avant la modélisation

Ces mesures permettent d'évaluer les données à dispositions dans le but de savoir si elles présentent des biais.

Déséquilibre des classes. Cette métrique permet de vérifier si les classes de la variable sensible sont équitablement représentées dans le jeu de données.

$$DC : \mathbb{P}(S = 1) - \mathbb{P}(S = 0).$$

Différence des proportions positives. Elle mesure si dans le jeu de données, les individus de certaines classes de S sont moins représentés dans les instances positives de Y.

$$\mathbb{P}(Y = 1|S = 1) - \mathbb{P}(Y = 1|S = 0).$$

+ Cas multiclasse : une approche peut être celle de calculer la différence de proportions pour chacune des valeurs de Y et d'analyser ces écarts. L'écart le plus grand peut être retenu.
+ Cas continu : calculer la différence de moyenne entre les valeurs réelles de Y pour chaque classe de S.

$$\mathbb{E}(Y|S = 1) - \mathbb{E}(Y|S = 0).$$

Divergence de Kullback–Leibler (KL). La divergence de KL permet de mesurer la différence entre deux distributions. Dans la théorie initialement introduite par Kullback et Leibler, l'idée était de comparer une distribution observée à une distribution théorique. Ainsi, pour deux distributions de probabilités P et Q, la divergence KL se calcule :

$$KL(P|Q) = \sum_x P(x)\log\left(\frac{P(x)}{Q(x)}\right).$$

Le résultat obtenu s'interprète comme étant la différence entre le nombre de bits nécessaire pour représenter des échantillons de P en utilisant des codes optimisés pour Q. Dans le cas présenté ici, une comparaison est effectuée entre la distribution empirique pour $S = 1$, $\mathbb{P}_1$, et celle pour $S = 0$, $\mathbb{P}_0$:

$$KL(\mathbb{P}_1|\mathbb{P}_0) = \sum_x \mathbb{P}_1(x)\log\left(\frac{\mathbb{P}_1(x)}{\mathbb{P}_0(x)}\right).$$

Divergence de Jensen–Shannon (JS). La divergence de JS est basée sur la divergence de KL. Elle se distingue par le fait qu'elle permet de mesurer les différences symétriques entre les deux distributions. Elle prend toujours une valeur finie positive. Cela permet de définir une mesure à partir de sa racine carrée. En posant, $\mathbb{P}_M$ la moyenne des distributions $\mathbb{P}_1$ et $\mathbb{P}_0$,

$$JS(\mathbb{P}_1, \mathbb{P}_0, \mathbb{P}_M) = \frac{1}{2}\left(KL(\mathbb{P}_1, \mathbb{P}_M) + KL(\mathbb{P}_0, \mathbb{P}_M)\right).$$

Les deux divergences présentées ci-dessus peuvent être utilisées pour mesurer les différences de distributions observées sur les variables explicatives comme sur la variable cible. Ces divergences sont adaptées à toutes formes de variables.

+ *Cas multiclasse :* Calcul des divergences entre les distributions $\mathbb{P}_1$ et $\mathbb{P}_0$ pour chacune des valeurs de Y. Le max pourrait être retenu.

+ *Cas continu :* Calcul des divergences entre les distributions continues :

$$KL(\mathbb{P}_1|\mathbb{P}_0) = \int_{-\infty}^{+\infty} \mathbb{P}_1(x) \log\left(\frac{\mathbb{P}_1(x)}{\mathbb{P}_0(x)}\right).$$

Disparité statistique conditionnelle. Cette mesure introduite en 2020 par Wachter et al.[78], permet de savoir si, conditionnellement à une variable choisie dans les données, une classe de la variable sensible est désavantagée par rapport à l'autre. Au lieu de regarder directement la liaison entre Y et S, cette liaison est considérée par rapport à une variable V choisie. Ce conditionnement permet d'éviter le paradoxe de Simpson. En effet, il a été mis en évidence dans de nombreux cas que certaines vérités s'inversaient en prenant en compte des cofacteurs associés au phénomène étudié.
Soit V la variable par rapport à laquelle est effectuée le conditionnement, V_i les différents sous-groupes de cette variable et n_i leur cardinal respectif. Pour chaque sous-groupe de V, les ratios suivants sont calculés :

$$D_i = \frac{\#\{Y = 0 | S = 1, V_i\}}{\#\{Y = 0\}}$$

$$A_i = \frac{\#\{Y = 1 | S = 1, V_i\}}{\#\{Y = 1\}}$$

Ainsi,

$$DSC = \frac{1}{n} \sum_{V_i} n_i DD_i,$$

où $DD_i = D_i - A_i$. Plus cette métrique est proche de 0 plus le modèle est équitable suivant le critère de parité statistique conditionnelle. Cette métrique mesure donc si en moyenne, dans les sous-groupes de V, les chances d'obtenir les différentes valeurs de Y

sont les mêmes quand $S = 1$. Cette métrique peut aussi être utilisée avec $\widehat{Y}$ à la place de Y. La Disparité statistique conditionnelle ne peut pas être définie dans le cas continu. Dans le cas de la classification multiclasse, une adaptation pourrait être de faire les calculs en opposant les classes deux à deux ou une contre les autres.

Distance de Kolmogorov-Smirnov. La distance de Kolmogorov-Smirnov est souvent utilisée pour comparer une distribution empirique à une distribution théorique. Cette distance s'adapte dans le cas où il faut comparer deux distributions empiriques entre elles. Une fois la distance calculée, elle peut être utilisée pour exécuter le test de Kolmogorov-Smirnov.

Dans le cas étudié ici, la distance se calcule de la manière suivante :

$$d_{KS} = \max(|\mathbb{P}_1 - \mathbb{P}_0|).$$

Ainsi, pour un niveau α donné, l'hypothèse nulle stipulant que les deux distributions sont issues de la même loi de probabilité est rejetée si :

$$d_{KS} > \sqrt{-\frac{1}{2} ln\left(\frac{\alpha}{2}\right) \times \frac{\#\{S=1\} + \#\{S=0\}}{\#\{S=1\} \times \#\{S=0\}}}.$$

Dans la même veine que le test de Kolmogorov-Smirnov, les tests statistiques permettant l'évaluation des relations entre deux variables peuvent aussi être utilisés (tests d'Anova, de Wilcoxon Mann-Whitney, des signes, de khi-deux, etc.).

La norme L_p. Les différences entre les distributions peuvent être calculées en considérant des normes. Ainsi, pour $p \geq 1$,

$$L_p(\mathbb{P}_1, \mathbb{P}_0) = \left(\sum_x |\mathbb{P}_1(x) - \mathbb{P}_0(x)|^p \right)^{\frac{1}{p}}.$$

Cette métrique ne peut pas prendre de valeur négative. Elle ne permet donc pas de distinguer les cas où le biais est contre $S = 1$ ou contre $S = 0$.

La distance en variation totale. Cette distance permet de mesurer la plus grande différence observable entre des probabilités allouées par les deux distributions pour des évènements identiques. Ainsi, en notant $\mathcal{F}$ la tribu sur laquelle est définie les probabilités,

$$d_{VT} = \sup_{A \in \mathcal{F}} |\mathbb{P}_1(A) - \mathbb{P}_0(A)|.$$

Pour des cas discrets comme celui considérés ici, cette distance peut se réécrire comme étant la moitié de la distance de L_1 :

$$d_{VT} = \frac{1}{2} L_1(\mathbb{P}_1, \mathbb{P}_0)$$

Les distances présentées ci-dessus peuvent être calculées sur tous types de distributions.

+ Cas multiclasse : calcul des distances entre les distributions $\mathbb{P}_1$ et $\mathbb{P}_0$ pour chacune des valeurs de Y. Le max pourrait être retenu.

+ Cas continu : calcul des différentes distances dans le cas continu.

2.5.3 Focus sur la régression

Le cas de la régression, beaucoup moins traité dans la littérature requiert une attention particulière. En effet, dans le cas binaire, Agarwal et al. ont démontré que les notions d'indépendance faible étaient suffisantes pour que les critères de parité statistique et d'égalité des chances soient vérifiés. Ainsi, dans le cas binaire :

$$\widehat{Y} \perp\!\!\!\perp S \Leftrightarrow \mathbb{E}(\widehat{Y}|S) = \mathbb{E}(\widehat{Y}) : \text{Parité statistique}^{[1]}.$$

$$\widehat{Y} \perp\!\!\!\perp S|Y \Leftrightarrow \mathbb{E}(\widehat{Y}|S,Y) = \mathbb{E}(\widehat{Y}|Y) : \text{Égalité des chances}^{[1]}.$$

Ces équivalences permettent d'utiliser certaines métriques moins contraignantes, mais toujours satisfaisantes du point de vue théorique. Ces équivalences ne sont plus valables dans le cas continu. Par exemple, la mesure d'impact disparate permet de vérifier que le critère de parité statistique est vérifié dans le cas binaire. Cependant, dans le cas continu, l'adaptation proposée ne permet pas de vérifier la définition au sens stricte du terme.

Les distances et les divergences introduites permettent de mieux traiter le cas continu. Quand Y est continue, une approche est de calculer des indicateurs de dépendance classique tels que corrélation Pearson, tau de Kendall et rho de Spearman. L'inconvénient de ces approches est qu'elles ne permettent de capter qu'un spectre limité des formes de dépendance. Des métriques plus complexes mesurant un plus large spectre de dépendance statistique sont donc introduites.

FairQuant. Cette métrique introduite par Grari et al. propose de mesurer l'équité d'un modèle en calculant des écarts absolus de prédiction entre d'une part la moyenne dans des sous-groupes construits à l'aide des quantiles de la variable sensible et, d'autre part, la moyenne globale.

Grari et al.[33] proposent de mesurer l'équité d'un modèle en subdivisant le jeu de données de test en K parties en se basant sur les quantiles de S. L'utilisation des quantiles permet d'obtenir des groupes de mêmes tailles. Dans chaque groupe, une moyenne des prédictions est calculée. Un écart absolu est ensuite calculé entre la moyenne observée sur toutes les données de test et les moyennes calculées par sous-groupes. En notant μ_i la moyenne dans le sous groupe i et μ la moyenne sur toutes les données de test :

Parité statistique :

$$FairQuant = \frac{1}{K}\sum_{i=1}^{K} |\mu_i - \mu|,$$

avec μ_i la moyenne des prédictions dans le sous ensemble i ; $\widehat{f}(X)$.

Chances égalisées :

$$FairQuant = \frac{1}{K}\sum_{i=1}^{K}|\mu_i - \mu|,$$

avec μ_i la moyenne des prédictions dans le sous ensemble i réduit de la valeur réelle de Y ; $\widehat{f}(X) - Y$.

HGR (Hirschfeld-Gebelein-Rényi). De nombreuses mesures ont été introduites et étudiées dans le but de capter le spectre de relations non linéaires le plus large possible. Des mesures telles que la covariance de distance brownienne[76], l'analyse des corrélations en utilisant des noyaux canoniques[37], le critère d'indépendance de Hilbert-Schmidt[35] etc. Ces différentes approches ont été critiquées pour des temps de calculs importants, des difficultés d'implémentation, des performances faibles en présence de bruit.

Le coefficient HGR (Hirschfeld-Gebelein-Rényi) introduit initialement en 1959[70] semble être un choix plus plébiscité dans la littérature. En effet, Rényi a défini sept propriétés fondamentales que doivent respecter une mesure de dépendance. Cet ensemble de propriétés est considéré comme une référence. Il a aussi prouvé que le coefficient HGR vérifie toutes ces propriétés. Les fondements théoriques de ce coefficient sont donc solides. De plus, il permet de mesurer de manière satisfaisante les relations linéaires et non linéaires, il est invariant aux modifications des distributions marginales et peut traiter les variables multidimensionnelles.

Pour deux variables aléatoires $Y \in \mathcal{Y}$ et $S \in \mathcal{S}$, le coefficient HGR se calcule de la manière suivante :

$$\begin{aligned} HGR(Y,S) &= \sup_{f:\mathcal{Y}\mapsto\mathbb{R},\ g:\mathcal{S}\mapsto\mathbb{R}} \rho(f(Y), g(S)) \\ &= \sup_{f:\mathcal{Y}\mapsto\mathbb{R},\ g:\mathcal{S}\mapsto\mathbb{R}} \mathbb{E}(f(Y), g(S)) \end{aligned} \tag{2.3}$$

Avec les conditions que $\mathbb{E}(f(Y)) = \mathbb{E}(g(Y)) = 0$ et $\mathbb{E}(f^2(Y)) = \mathbb{E}(g^2(Y)) = 1$.
Les fonctions f et g étant choisies parmi une infinité de fonctions, le calcul de ce coefficient s'avère difficile, voire impossible, sous ce format strict. Il s'agit plus d'un concept abstrait que d'une réelle mesure de dépendance. Toutefois, de nombreuses adaptations conservant la même structure ont été proposées dans le but d'approcher la vraie valeur de HGR. Ces adaptations ont prouvé leur performance dans de nombreux cas d'usage. Ci-dessous sont présentées quatre de ces adaptations.

Randomized Dependence Coefficient (RDC). Cette méthode mesure la dépendance entre deux variables aléatoires comme étant la plus grande corrélation canonique entre k projections non linéaires des transformations de leur copule. La figure 2.15 illustre les étapes de calcul du RDC. Cette illustration provient des travaux Lopez-Paz et al.[51]

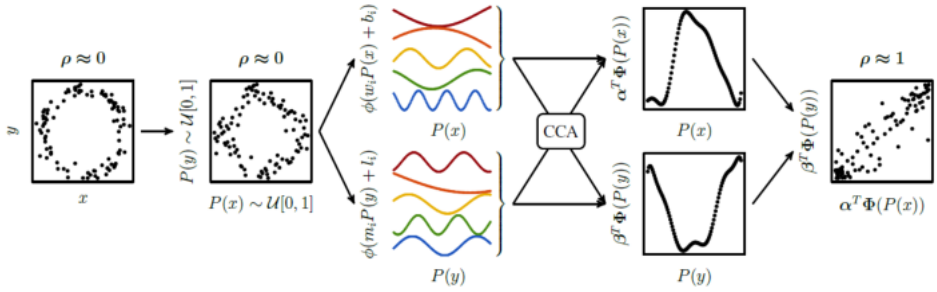


FIGURE 2.15 – Illustration des étapes de calcul du RDC.

sur le sujet. CCA représente l'analyse de corrélation canonique (Canonical Correlation Analysis).

Ce coefficient simple est à implémenter et à des temps d'exécution raisonnables. Il conserve aussi de bonnes propriétés. En effet, il est invariant aux modifications des distributions marginales et respecte, suivant une certaine marge de prudence, les sept propriétés fondamentales.

HGR estimé à l'aide de méthodes d'estimation par noyau. En 1975, Witsenhausen introduit une caractérisation du coefficient HGR en utilisant la seconde valeur singulière d'une matrice construite ingénieusement[80]. Cette caractérisation permet de construire une majoration du carré du HGR. En partant de ces travaux, Mary et al.[55], utilisent des méthodes d'estimation par noyau dans le but d'estimer cette borne supérieur. Ils prouvent empiriquement que cette approche fournit de bons résultats et cela même en présence de bruits.

HGR estimé en utilisant des réseaux neuronaux. La principale limite de l'approche précédente réside dans le choix du noyau. Une fois le noyau choisi, il faut aussi le paramétrer. Ce coefficient peut donc être difficile à généraliser à tous types de jeu de données. Une solution proposée par Grari et al. est de remplacer la méthode d'estimation par noyau par un réseau de neurones[33]. Les fonctions f et g du HGR sont remplacées par des réseaux de neurones. Ces réseaux sont entraînés sur des échantillons issus de la distribution jointe des deux variables à étudier. À chaque itération, les sorties de ces réseaux sont standardisées. Le coefficient HGR est estimé comme étant la moyenne des produits des sorties des réseaux.
L'utilisation des réseaux donne plus de flexibilité à cette approche et permet de couvrir un plus large spectre de transformations. Toutefois, cette méthode peut s'avérer coûteuse en temps de calcul et plus difficile à paramétrer du fait du choix de la structure optimale du réseau.

Alternative Conditional Expectations (ACE). Breiman et Friedman introduisent, en 1985[9], une méthode non paramétrique permettant de trouver les bonnes fonctions à utiliser pour transformer les variables d'une régression linéaire multiple et ainsi prendre

en compte les relations non linéaires présentent. Pour se faire, ils utilisent la méthode de la puissance itérée.
De par sa définition, cette méthode peut servir à mesurer la dépendance entre deux variables. En effet, il suffit de chercher la transformation optimale, par exemple celle qui maximise le R^2 et de retenir le maximum des R^2 comme étant une mesure de dépendance. En se basant sur les fondements solides développés par ces auteurs, il est possible de prouver que cette approche vérifie les sept propriétés fondamentales de Rényi. Cependant, elle est difficile à implémenter et peut perdre en performance en présence de bruits additifs.

Toutes les approches fournies dans cette section permettent de mesurer plus ou moins l'équité des données et des modèles. Il est indispensable de faire un choix critique des métriques à utiliser en prenant en compte le cas d'usage, les fondements théoriques, le spectre d'équité couvert et l'efficacité opérationnelle (implémentation et temps de calcul).

2.6 Incompatibilité des mesures de biais de groupe

Les discussions menées dans la section mesure individuelle et mesure de groupe, ont mis en avant les points de divergence entre ces deux groupes de critères. Ces deux types d'équité étant difficiles à atteindre simultanément. En ce qui concerne les critères individuels, ils sont tous définis sur la base de la notion de similarité. Les critères varient donc suivant leur méthode de traitement de cette similarité, mais il s'agit bien du même spectre d'équité.

Les critères d'équité de groupe peuvent, quant à eux être, regroupés en trois grands familles. Ces familles représentent des visions différentes de la notion d'équité de groupe. A première vue, ces différences ne semblent pas significatives. Il peut sembler qu'elles relèvent plus d'une gymnastique mathématique. Cette vision est pourtant erronée. En effet, sauf dans des cas dégénérés, ces trois définitions de l'équité de groupe sont incompatibles. En d'autres termes, l'atteinte d'un de ces objectifs empêche l'atteinte des autres.

2.6.1 Chances égalisées et parité statistique

Les chances égalisées et la parité statistique ne peuvent être atteintes en même temps que dans le cas où $S \perp\!\!\!\perp Y$ ou $\widehat{Y} \perp\!\!\!\perp Y$.
En effet, en supposant une dépendance entre S et Y ou entre $\widehat{Y}$ et Y, les chances égalisées et la parité statistique ne peuvent être vérifiées simultanément.

Justifications : $S \not\perp\!\!\!\perp Y$ ou $\widehat{Y} \not\perp\!\!\!\perp Y$ n'empêche pas $\widehat{Y} \perp\!\!\!\perp S|Y$ (chances égalisées) et n'empêche pas $S \perp\!\!\!\perp \widehat{Y}$ (parité statistique). Toutefois, dans le cas où ces deux critères sont vérifiés simultanément :

$$\mathbb{P}(\widehat{Y}|Y,S) = \underbrace{\mathbb{P}(\widehat{Y}|S)}_{\text{Egalité des chances}} = \underbrace{\mathbb{P}(\widehat{Y})}_{\text{Parité statistique}}.$$

Ainsi, $\widehat{Y} \perp\!\!\!\perp Y$ ce qui est contradictoire avec le fait que Y et $\widehat{Y}$ soient dépendantes.

En règle générale, la variable S est corrélée (à un effet significatif qu'il soit direct ou indirect) avec la réponse Y. Si ce n'est pas le cas, l'application de la théorie de mitigation du biais perd une grande partie de son intérêt. En effet, si $S \perp\!\!\!\perp Y$, il n'y a pas de biais présent dans les données initiales. Le seul biais qui puisse exister est un biais introduit dans les modèles.
De plus, $\widehat{Y}$ et Y sont dépendantes. En effet, $\widehat{Y}$ est construite dans le but d'être la plus proche possible de Y. Si elles sont indépendantes, le modèle construit n'a aucun intérêt. Ce sont donc tous les deux des cas dégénérés.

Une preuve dans le cas binaire est la suivante. En utilisant le théorème de probabilité totale :

$$\begin{aligned}\mathbb{P}(\widehat{Y}=1|S=s) &= \mathbb{P}(Y=1|S=s)\mathbb{P}(\widehat{Y}=1|Y=1,S=s) + \mathbb{P}(Y=0|S=s)\mathbb{P}(\widehat{Y}=1|Y=0,S=s)\\ &= \mathbb{P}(Y=1|S=s)TVP_s + \mathbb{P}(Y=0|S=s)TFP_s.\end{aligned}$$

La différence suivante est calculée :

$$\begin{aligned}\mathbb{P}(\widehat{Y}=1|S=0) - \mathbb{P}(\widehat{Y}=1|S=1) &= \mathbb{P}(Y=1|S=0)TVP_0 + \mathbb{P}(Y=0|S=0)TFP_0\\ &\quad - \mathbb{P}(Y=1|S=1)TVP_1 - \mathbb{P}(Y=0|S=1)TFP_1.\end{aligned}$$

En supposant la condition de chances égalisées vérifiée :

$$TPV_0 = TVP_1 = \mathbb{P}(\widehat{Y}=1|Y=1)$$

et

$$TFP_0 = TFP_1 = \mathbb{P}(\widehat{Y}=1|Y=0).$$

L'égalité précédente devient :

$$\begin{aligned}&\mathbb{P}(\widehat{Y}=1|S=0) - \mathbb{P}(\widehat{Y}=1|S=1)\\ &\quad= \mathbb{P}(\widehat{Y}=1|Y=1)(\mathbb{P}(Y=1|S=0) - \mathbb{P}(Y=1|S=1))\\ &\quad+ \mathbb{P}(\widehat{Y}=1|Y=0)(\mathbb{P}(Y=0|S=0) - \mathbb{P}(Y=0|S=1))\\ &\quad= \{\mathbb{P}(Y=1|S=0) - \mathbb{P}(Y=1|S=1)\}\{\mathbb{P}(\widehat{Y}=1|Y=1) - \mathbb{P}(\widehat{Y}=1|Y=0)\}.\end{aligned}$$

Pour que la parité statistique soit vérifiée, il faut que $\mathbb{P}(\widehat{Y}=1|S=0) - \mathbb{P}(\widehat{Y}=1|S=1)$ soit nulle. Ceci implique que le terme de droite est nul. Ainsi, soit $\mathbb{P}(Y=1|S=0) = \mathbb{P}(Y=1|S=1)$ ou soit $\mathbb{P}(\widehat{Y}=1|Y=1) = \mathbb{P}(\widehat{Y}=1|Y=0)$. Ces possibilités conduisent aux conclusions suivantes :

— $\mathbb{P}(Y = 1|S = 0) = \mathbb{P}(Y = 1|S = 1)$: cela signifie que quel que soit s, la chance d'obtenir la prédiction favorable est la même. Ainsi, l'effet discriminatoire, s'il existe, est faible.
— $\mathbb{P}(\widehat{Y} = 1|Y = 1) = \mathbb{P}(\widehat{Y} = 1|Y = 0)$: les taux de vrais positifs et de faux positifs sont égaux. Cela signifie que le modèle construit ne parvient pas à classer les individus de la classe $Y = 1$. En pratique, cela n'a aucun intérêt. En effet, le modèle construit doit pouvoir présenter un taux de vrais positifs significativement plus élevé que le taux de faux positifs.

Ces deux cas étant dégénérés, les deux critères ne sont donc atteignables que dans des cas dégénérés.

2.6.2 Parité prédictive et parité statistique

Les parités prédictive et statistique ne sont vérifiées simultanément que quand S est indépendante de Y.

Justifications :

$$\mathbb{P}(S|\widehat{Y}, Y) = \underbrace{\mathbb{P}(S|\widehat{Y})}_{\text{Parité prédictive}} = \underbrace{\mathbb{P}(S).}_{\text{Parité statistique}}$$

Cela est contradictoire avec le fait que S et Y soient dépendantes. Comme discuté plus haut, ce cas d'indépendance est dégénéré.

Dans le cas binaire, en supposant la condition de parité statistique vérifiée, c'est-à-dire $\mathbb{P}(\widehat{Y} = 1|S = 1) = \mathbb{P}(\widehat{Y} = 1|S = 0) = \mathbb{P}(\widehat{Y} = 1)$ et en utilisant le théorème de Bayes :

$$\begin{aligned} VPP_0 - VPP_1 &= \frac{TVP_0\ \mathbb{P}(Y = 1|S = 0)}{\mathbb{P}(\widehat{Y} = 1|S = 0)} - \frac{TVP_1\ \mathbb{P}(Y = 1|S = 1)}{\mathbb{P}(\widehat{Y} = 1|S = 1)} \\ &= \frac{TVP_0\ \mathbb{P}(Y = 1|S = 0) - TVP_1\ \mathbb{P}(Y = 1|S = 1)}{\mathbb{P}(\widehat{Y} = 1)} \end{aligned} \tag{2.4}$$

La vérification de la condition de parité prédictive requiert que $VPP_0 - VPP_1 = 0$. Cela revient à $TVP_0\ \mathbb{P}(Y = 1|S = 0) = TVP_1\ \mathbb{P}(Y = 1|S = 1)$, d'où :

$$\frac{TVP_0}{TVP_1} = \frac{\mathbb{P}(Y = 1|S = 1)}{\mathbb{P}(Y = 1|S = 0)}.$$

Ainsi, ces deux conditions peuvent être satisfaites concomitamment. Néanmoins, si les ratios sont significativement différents de 1, le modèle obtenu perd son utilité. En effet, le taux de vrai positif d'un des groupes devra être faible pour pouvoir assurer l'égalité. Par ailleurs, si ces ratios sont proches de 1, $\mathbb{P}(Y = 1|S = 1) \approx \mathbb{P}(Y = 1|S = 0)$. Comme expliqué plus haut, ce cas laisse peu de place à l'existence du biais dans les données.

2.6.3 Parité prédictive et chances égalisées

La parité prédictive et les chances égalisées ne sont vérifiées simultanément que quand S est indépendante de Y.

Justifications :
— $\widehat{Y} \perp\!\!\!\perp S|Y \implies S \perp\!\!\!\perp \widehat{Y}|Y$ (égalité des chances)
— $Y \perp\!\!\!\perp S|\widehat{Y} \implies S \perp\!\!\!\perp Y|\widehat{Y}$ (parité prédictive)
D'où $S \perp\!\!\!\perp (\widehat{Y}, Y) \implies S \perp\!\!\!\perp Y$. Cela est contradictoire avec le fait que S et Y soient dépendantes. Comme discuté plus haut, ce cas d'indépendance est dégénéré.

Dans le cas binaire, les deux conditions sont vérifiées si :

$$TVP_0 = TVP_1,\ TFP_0 = TFP_1 \text{ et } VPP_0 = VPP_1.$$

Ainsi,

$$\begin{aligned}\mathbb{P}(\widehat{Y}=1|S) &= \mathbb{P}(\widehat{Y}=1|Y=1,S)\mathbb{P}(\widehat{Y}=1|S) + \mathbb{P}(\widehat{Y}=1|Y=0,S)\mathbb{P}(\widehat{Y}=1|S) \\ &= TVP_0\ \mathbb{P}(Y=1|S) + TFP_0\ \mathbb{P}(Y=0|S).\end{aligned} \tag{2.5}$$

Dans le cas $S=0$,

$$\begin{aligned}TVP_0\ \mathbb{P}(Y=1|S=0) &= \mathbb{P}(\widehat{Y}=1|Y=1,S=0)\mathbb{P}(Y=1|S=0) \\ &= \frac{\mathbb{P}(Y=1|\widehat{Y}=1,S=0)\mathbb{P}(\widehat{Y}=1|S=0)}{\mathbb{P}(Y=1|S=0)}\mathbb{P}(Y=1|S=0) \\ &= \mathbb{P}(Y=1|\widehat{Y}=1,S=0)\left\{TVP_0\ \mathbb{P}(\widehat{Y}=1|S=0) + TFP_0\ \mathbb{P}(Y=0|S=0)\right\} \\ &= VPP_0\left\{TVP_0\ \mathbb{P}(\widehat{Y}=1|S=0) + TFP_0\ (1-\mathbb{P}(Y=1|S=0)\right\}.\end{aligned} \tag{2.6}$$

D'où,

$$\mathbb{P}(Y=1|S=0) = \frac{VPP_0\ TFP_0}{VPP_0\ TFP_0 + (1-VPP_0)TVP_0}.$$

De même, dans le cas $S=1$,

$$\mathbb{P}(Y=1|S=1) = \frac{VPP_1\ TFP_1}{VPP_1\ TFP_1 + (1-VPP_1)TVP_1}.$$

Ainsi, pour que nos deux conditions soient vérifiées, il faut que $\mathbb{P}(Y=1|S=1) = \mathbb{P}(Y=1|S=0)$. Cette égalité conduit à un cas dans lequel l'étude de la discrimination perd de son utilité puisque le traitement entre les deux modalités de la variable sensible est identique dans le jeu de données. Dans le cas où le modèle est parfait en prédiction,

l'équation devient une forme indéterminée 0/0. Il n'est donc pas possible de tirer des conclusions sur les différents taux observés.

Il est intéressant de remarquer que la condition d'égalité des opportunités est compatible avec la parité prédictive. En effet, l'égalité des opportunités est une variante moins contraignante des chances égalisées (ne requiert que l'égalité des taux de vrais positifs). En retirant la condition $TFP_0 = TFP_1$, les deux conditions peuvent être vérifiées simultanément. Toutefois, si $\mathbb{P}(Y = 1|S = 1)$ est significativement différent de $\mathbb{P}(Y = 1|S = 0)$, un des taux de faux positifs sera significativement plus grand que l'autre. Ce qui peut être très préjudiciable dans certaines applications.

L'incompatibilité de ces différents critères met en avant l'utilité de présenter de les subtilités des différentes notions de biais. Cette compréhension globale du sujet et de ses différentes nuances permet d'interpréter avec plus de recul les résultats qui sont obtenus par les métriques. Au-delà de l'incompatibilité des différentes notions, le coût de leur implémentation est un sujet important. Intuitivement, l'ajout de nouvelles contraintes devrait conduire à la réduction de la performance globale des modèles. La section suivante permet d'approcher de manière mathématique le coût qui pourrait être observé en pratique. En plus de la performance, il peut aussi arriver que les temps de calculs deviennent plus long.

2.7 Le coût de l'équité

La construction de modèle plus juste à un coût. De nouvelles contraintes sont intégrées dans le processus de modélisation ce qui peut avoir pour conséquence de réduire la qualité des modèles obtenus. Cette partie fournit une vision globale de la quantification de cette perte de précision dans le cas de la parité statistique et des chances égalisées. Comme introduit dans la section 3.1, l'objectif est de trouver une fonction de la famille $\mathbb{F}$ qui permet de minimiser l'erreur L. Les contraintes d'équité nécessite la définition de nouvelles classes :

$$\mathbb{F}_{PS} = \{f(X,S) \in \mathbb{F} \ / \ \widehat{Y} \perp\!\!\!\perp S\} \qquad \text{(Parité statistique)}$$
$$\mathbb{F}_{EC} = \{f(X,S) \in \mathbb{F} \ / \ \widehat{Y}|Y \perp\!\!\!\perp S\} \qquad \text{(Chances égalisées)}$$

En notant $R(f) = \mathbb{E}(L(Y, f(X,S)))$ la performance d'un modèle, il est possible de définir le coût de l'équité comme étant :

$$\widehat{\Delta}_c(\mathbb{F}) = \inf_{f \in \mathbb{F}_{EC} \text{ ou } \mathbb{F}_{PS}} R(f) - \inf_{f \in \mathbb{F}} R(f).$$

2.7.1 Parité statistique

Régression. La condition de parité statistique peut se réécrire en utilisant des distributions. La condition devient :

$$\mathcal{L}(f(X,S)|S) = \mathcal{L}(f(X,S)).$$

Il faudrait donc que $\forall s$ et $\forall A$ ensemble mesurable, $\mathbb{P}(f(X,S) \in A|S=s) = \mathbb{P}(f(X,S) \in A)$. Dans le cadre de la régression, la fonction de performance peut prendre la forme suivante :

$$R(\mathbb{F}) = \min_{f \in \mathbb{F}} \mathbb{E}[||Y - f(X,S)||^2].$$

Avec les données à disposition, le meilleur estimateur est $f^* = \mathbb{E}[Y|(X,S)]$. Le Gouic et al.[32] définissent cet estimateur comme étant l'estimateur de Bayes qui minimise le risque bayésien défini plus haut.
En notant, μ_s la distribution conditionnelle de l'estimateur sachant S et $\nu_S(f)$ la distribution conditionnelle de $f(X,S)$ sachant S, il est possible d'utiliser la distance de Wasserstein pour obtenir une borne inférieure. Pour deux mesures de probabilité P et Q sur $\mathbb{R}^d$ cette distance se définit comme étant :

$$W_2^2(P,Q) = \min_{\lambda \in \Lambda(P,Q)} \int ||x-y||^2 d\lambda(x,y),$$

où $\Lambda(P,Q)$ est l'ensemble de toutes les mesures de probabilités sur $\mathbb{R}^d \times \mathbb{R}^d$ ayant pour marginales P et Q.

Théorème

$$\widehat{\Delta}_c(\mathbb{F}) \geq \inf_{g \in \mathbb{F}} \mathbb{E}[W_2^2(\mu_S, \nu_S(g))].$$

De plus, si $\mathbb{F} = \mathbb{F}_{PS}$ et $\forall s,\ \mu_s$ a une densité par rapport à la mesure de Lebesgue, alors la relation devient une égalité. Dans le cas de la parité statistique, la fonction de prédiction construite est indépendante de S par définition. Ainsi, la relation se réécrit :

$$\inf_{g \in \mathbb{F}} \mathbb{E}[W_2^2(\mu_S, \nu_S(g))] = \inf_{\nu(g)} \mathbb{E}[W_2^2(\mu_S, \nu_S(g))].$$

La résolution de ce problème revient à la minimisation de :

$$\nu \mapsto \mathbb{E}[W_2^2(\mu_S, \nu)].$$

Ce problème revient à trouver le barycentre de Wasserstein. En effet, dans la littérature[3], il a été prouvé qu'en prenant $\mathbb{P}_S$ l'ensemble des distributions aléatoires μ_S, le minimum était atteint par le barycentre de Wasserstein de $\mathbb{P}_S$. Ainsi, la contrainte d'équité imposée à un prix non nul sur la performance du modèle construit. La mesure de ce prix se fait à travers la minimisation de la distance de Wasserstein.

Il est intéressant de remarquer que des formulations plus faibles de la parité statistique ont été étudiées dans la littérature. Ainsi, Dwork et al.[21], définissent la parité statistique en utilisant des contraintes d'indépendance plus faibles : $\mathbb{E}[f(X,S)|S] = \mathbb{E}[f(X,S)]$ ou $\text{Cov}(f(X,S)) = 0$. Le but étant d'avoir des contraintes moins coûteuses en performance et en temps de calcul. Dans le cas où S est une variable binaire (le cas du genre par exemple $S \in \{0,1\}$), les critères se réécrivent :

$$\begin{aligned} \mathbb{E}_{X,S}[f(X,S)] &= \mathbb{E}_X[\mathbb{E}_S[f(X,S)|S]] \\ &= \mathbb{P}(S=0)\mathbb{E}_X[f(X,S)|S=0] + \mathbb{P}(S=1)\mathbb{E}_X[f(X,S)|S=1]. \end{aligned} \tag{2.7}$$

L'expression de l'espérance conditionnelle peut se réécrire :

$$\begin{aligned}\mathbb{E}_{X,S}[f(X,S)|S] &= \frac{\mathbb{E}_{X,S}[Sf(X,S)]}{\mathbb{E}[S]} \\ &= \frac{\mathbb{P}(S=1)\mathbb{E}_X[f(X,S)|S=1]}{\mathbb{P}(S=1)} \\ &= \mathbb{E}_X[f(X,S)|S=1].\end{aligned} \tag{2.8}$$

Ainsi, pour $\mathbb{P}(S=1)>0$, la condition de parité statistique est vérifiée si et seulement si

$$\mathbb{E}_X[f(X,S)|S=0] = \mathbb{E}_X[f(X,S)|S=1].$$

Classification. Dans le cadre général de classification, il n'existe pas encore de théorème permettant de quantifier ou de borner la perte de performance. Toutefois, certains résultats ont pu être prouvés dans des cadres plus restreints. Jiang et al.[42] ont prouvé que dans le cadre d'un modèle de classification utilisant des seuils de décisions, il était possible de réduire le biais en minimisant le nombre de changement de classes induit par la prise en compte de la contrainte. Les auteurs assurent le respect de la contrainte en minimisant la distance de Wasserstein d'ordre 1 entre les individus ayant des modalités de la variable sensible différentes. Ils ont prouvé l'existence d'une borne pour $\widehat{\Delta}_c$ qui est atteinte pour le barycentre de Wasserstein d'ordre 1.

D'autres part, en se restreignant aux modèles entraînés par une transformation des données, il a été prouvé qu'il est possible de majorer la perte de performance[31]. Ces méthodes de transformations de données utilisent les distributions conditionnelles dans le but de faire disparaître toutes relations entre la variable sensible et les autres variables. Il s'agit du nettoyage des données. Une fois les données transformées, l'effet de l'utilisation de la transformation sur la performance est évalué à l'aide de la distance de Wasserstein. Comme dans les cas étudiés précédemment, la valeur optimale est obtenue en calculant le barycentre de Wasserstein.

2.7.2 Chances égalisées

Régression. La définition des chances égalisées est très contraignante. En pratique, il est difficile, voire impossible, de l'atteindre parfaitement. Dans "Learning nondiscriminatory predictors", les auteurs[81] montrent que du point de vue informatique, le temps de calcul est trop long. Par exemple, pour mettre en perspective le coût calculatoire, les auteurs se placent dans le cadre des fonctions linéaires, avec des fonctions de perte convexes et en imposant que seul le signe de la fonction de prédiction soit équitable d'un groupe à l'autre. Dans ce cas simplifié, l'application de la condition d'équité des chances requiert un temps de calcul exponentiel dans le pire des cas. Des solutions pourraient être de se tourner vers des critères moins contraignants comme l'égalité des opportunités, ou de choisir des mesures d'"indépendance" moins strictes telle que la corrélation. Le cas gaussien linéaire est un des cas pour lesquels les calculs sont envisageables suivant la définition stricte.

Considérant le cadre linéaire gaussien standard, avec $X \in \mathbb{R}^{n\times p}$, $S \in \mathbb{R}^{n\times 1}$ et $(\epsilon)_{1\geq i\geq n} \sim \mathcal{N}(0,1)$ les résidus i.i.d., Y s'écrit :

$$Y = f_{\alpha,\beta}(X,S) + \epsilon,$$

avec $f_{\alpha,\beta}(X,S) = \beta^t X + \alpha S$ une fonction linéaire, $\alpha \in \mathbb{R}$ et $\beta \in \mathbb{R}^{p\times 1}$. La distribution jointe de (X,S,Y) s'écrit :

$$(X,S,Y) \sim \mathcal{N}\left\{ \begin{pmatrix} \mu_X \\ \mu_S \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_X & \Sigma_{XS} & \Sigma_{XY} \\ \Sigma_{XS}^T & \Sigma_S & \Sigma_{ST} \\ \Sigma_{XY}^T & \Sigma_{SY}^T & \Sigma_Y \end{pmatrix} \right\}.$$

Dans ce cadre, la condition à vérifier $f_{\alpha,\beta}(X,S) \perp\!\!\!\perp S|Y$ est équivalente à une version relâchée

$$Cov(f(X,S), S|Y) = 0.$$

Cette équivalence est ce qui permet de calculer l'estimateur sous la condition stricte des chances égalisées. Cette relaxation est souvent envisagée dans les cas non gaussien et linéaire. Ainsi, trouver un estimateur qui respecte cette contrainte revient à résoudre :

$$(\widehat{\alpha},\widehat{\beta}) = \underset{(\alpha,\beta)\in\mathbb{F}_{EC}}{\arg\min}\ \mathbb{E}\left[Y - f_{\alpha,\beta}(X,S))^2\right],$$

où $\mathbb{F}_{EC}$ se réécrit :

$$\{(\alpha,\beta) \in \mathbb{R}\times\mathbb{R}^p \ \ / \ \ \beta^t(\Sigma_{XS}\Sigma_Y - \Sigma_{SY}\Sigma_{XY}) + \alpha(\Sigma_S\Sigma_Y - \Sigma_{SY}^2) = 0\}.$$

Cette écriture provient du vecteur de correction pour équité :

$$C_{S,X,Y} := \left(\frac{\Sigma_{XS}\Sigma_Y - \Sigma_{SY}\Sigma_{XY}}{\Sigma_S\Sigma_Y - \Sigma_{SY}^2}\right).$$

La solution exacte du programme de minimisation est :

$$\widehat{\alpha} = \widehat{\beta}^t C_{S,X,Y},$$

$$\widehat{\beta} = (\Sigma_X + \Sigma_S C_{S,X,Y} C_{S,X,Y}^t + C_{S,X,Y}\Sigma_{XS}^t + \Sigma_{XS} C_{S,X,Y}^t)^{-1}(\Sigma_{XY} + \Sigma_{SY} C_{S,X,Y}).$$

Cette solution exacte existe quand Y et S ne sont pas linéairement dépendantes.

Classification. Dans le but d'être en capacité d'obtenir des résultats quant à la quantification du coût de l'équité dans le cas de la classification, les hypothèses suivantes sont formulées :
$Y \in \{0,1\}$ et $S \in \{0,1\}$ telles que toutes les classes et les combinaisons de classes entre Y et S soient représentées. Ainsi, $\mathbb{P}(Y=0) \notin \{0,1\}, \mathbb{P}(S=0) \notin \{0,1\}$ et $\mathbb{P}(Y=0, S=0) \notin \{0,1\}$.

De plus, le modèle de classification se définit comme étant une règle de seuil pour une fonction de régression donnée. La solution f recherchée est donc de la forme : $1_{\{r(X,S)\geq \text{seuil}\}}$. r étant la fonction de régression,

$$r(x,s) = \mathbb{P}(Y=1|X=x,S=s) = \mathbb{E}[Y|X=x,S=s].$$

Cette fonction doit être telle que $\forall s \in \{0,1\}$, $t \mapsto \mathbb{P}(r(X,S) \leq t|S=s)$ soit continue sur $]0,1[$. Sous ces différentes hypothèses, le modèle optimal répondant à la problématique de minimisation :

$$\underset{f\in\mathbb{F}_{EC}}{\arg\min}\, \mathbb{E}(L(Y,f(X,S)))$$

où $\mathbb{F}_{EC}$ se réécrit :

$$\{f \in \mathbb{F} \ / \ \mathbb{P}(g(X,S)=i|Y=i,S=0) = \mathbb{P}(g(X,S)=i|Y=i,S=1), i=0,1\}$$

se définit comme étant le modèle $\widehat{g}$ tel que :

$$\widehat{g}(x,1) = 1_{\left\{1\leq 2r(X,1)-\widehat{\lambda}_1 \frac{r(X,1)}{\mathbb{P}(Y=1,S=1)}+\widehat{\lambda}_0 \frac{1-r(X,1)}{\mathbb{P}(Y=0,S=1)}\right\}},$$

où $(\widehat{\lambda}_0, \widehat{\lambda}_0) \in \mathbb{R}^2$ sont obtenues en résolvant les équations suivantes :

$$\frac{\mathbb{E}_{X|S=1}[r(X,1)\widehat{f}(X,1)]}{\mathbb{P}(Y=1|S=1)} = \frac{\mathbb{E}_{X|S=0}[r(X,0)\widehat{f}(X,0)]}{\mathbb{P}(Y=1|S=0)}$$

$$\frac{\mathbb{E}_{X|S=1}[(1-r(X,1))\widehat{f}(X,1)]}{\mathbb{P}(Y=0|S=1)} = \frac{\mathbb{E}_{X|S=0}[(1-r(X,0))\widehat{f}(X,0)]}{\mathbb{P}(Y=0|S=0)}.$$

Il s'agit du seul cas dans lequel une solution exacte peut être approchées.

Les éléments définis plus haut permettent de dire que l'équité a, en règle générale, un coût non nul sur la performance des modèles et sur les temps de calculs. Ainsi, dans la recherche de l'équité, il faut avoir recours à des méthodes permettant de préserver la performance des modèles. Le but est d'avoir une fonction prédictive permettant de maximiser la performance tout en minimisant le biais. Ces deux contraintes n'étant pas forcement compatible, il est nécessaire de se fixer des objectifs suivant le cas d'usage. Il faudra soit choisir un niveau maximum de perte de performance acceptable par rapport au modèle non contraint et maximiser l'équité du modèle, soit choisir le niveau minimum d'équité acceptable et maximiser la performance du modèle.

Chapitre 3

Mitigation du biais

Dans le chapitre précédent, les éléments permettant la mesure du biais ont été abordés. Une fois le biais défini et mesuré, le sujet de sa mitigation peut être traité. Il existe trois grandes familles d'approches qui sont développées par la suite :

1. Mitigation ante modélisation, par le retraitement des données ;
2. Mitigation pendant la modélisation, par ajustement des algorithmes ;
3. Mitigation post modélisation, par ajustement des prédictions.

Le type de mitigation se définit donc par l'étape à laquelle intervient cette mitigation dans le processus de modélisation. La figure 3.1 offre un résumé des éléments abordés dans les pages qui suivent.

FIGURE 3.1 – Les différentes approches de mitigation du biais.

3.1 Mitigation par le retraitement des données ante modélisation

La mitigation ante modélisation semble être la plus simple à aborder. Elle revient à transformer les données pour faire disparaître tout biais. Cette approche permet de

conserver ensuite toute la chaîne de modélisation initiale. Elle fait toutefois l'hypothèse que le reste de la chaîne de mitigation n'est source d'aucun biais. En effet, mitiger le biais avant la modélisation pour ensuite utiliser des algorithmes ou règles de décision biaisées ne résout pas vraiment le problème. Une solution est de mesurer les effets de cette mitigation avant la modélisation mais aussi après pour s'assurer que la modélisation n'amplifie pas le biais. Un autre des inconvénients de cette approche est le risque de perte d'informations. Aucune des méthodes de pré-traitement ne peut prétendre pouvoir extraire le biais des données comme un chirurgien qui opère méticuleusement son patient. De plus, le chirurgien, même s'il réussit à ne pas commettre d'erreur, laisse après son passage des cicatrices. Toute cette analogie pour dire qu'en retirant le biais des données, il est probable que des informations pertinentes, non biaisées soient perdues.

Ainsi, les méthodes de cette section doivent réduire le biais tout en conservant le maximum d'informations. Ces méthodes s'apparentent donc théoriquement à des méthodes dites de représentation, c'est-à-dire transformer X en un $\tilde{X}$ non biaisé en minimisant la distance entre X et $\tilde{X}$.

3.1.1 Suppression totale

L'approche la plus intuitive à cette mitigation est la suppression de la variable sensible et de toutes les variables qui lui sont liées. La suppression de la variable sensible permet de traiter la discrimination directe et la suppression des variables dépendantes permet de traiter la discrimination indirecte. Toutefois, il est clair que la quantité d'informations perdue est probablement significative. Pour que les performances soient maintenues, il faudrait que les variables décorrélées de la variable sensible soient suffisamment importantes pour compenser la perte d'informations engendrée par la suppression.
Ainsi, en partant de $(X_1, X_2, \ldots, X_n)$, n variables explicatives et S la variable sensible,

- La dépendance entre S et chacune des X_i est mesurée. Cette mesure de dépendance peut être choisie parmi les mesures présentées plus haut. Le coefficient HGR basé sur les noyaux est choisi dans les applications à cause de ses propriétés théoriques.
- Un seuil de dépendance maximal entre S et X_i est choisi.
- Toutes les variables ayant un niveau de dépendance supérieur au seuil sont supprimées de la base de données qui sera ensuite utilisée pour la modélisation.

Le choix du seuil peut se faire en effectuant un trade off entre performance et biais mitigé en construisant, par exemple, des courbes qui retracent l'évolution de ces éléments par seuil. Il faudra ensuite choisir soit un niveau de biais maximum acceptable ou un niveau de performance minimum.

3.1.2 Suppression de corrélation linéaire

Au lieu de supprimer toutes les variables liées à la variable sensible S, il est possible d'envisager une transformation des variables permettant la réduction des interdépendances tout en conservant les structures présentes dans les données. La méthode

de suppression de corrélation propose d'atteindre cet objectif en résolvant un programme de minimisation sous contraintes.

$$\min_{X_1,X_2,\ldots,X_n} \sum_{i=1}^{n} ||(x_1^i,\ldots,x_n^i)^T - (x_1^i,\ldots,x_n^i,s^i)^T||^2,$$

sous contrainte que :

$$\frac{1}{n}\sum_{i=1}^{n}(x_1^i,\ldots,x_n^i)(s^i-\bar{s})^T = 0.$$

Pour résoudre ce problème, un modèle linéaire expliquant les observations X à partir de la variable sensible centrée est construit :

$$X = \beta(S-\bar{S}) + \epsilon.$$

La nouvelle valeur $\tilde{X}$ retraitée de X est obtenue en récupérant les résidus de cette régression :

$$\tilde{X} = X - \beta(S-\bar{S}).$$

En effet, l'idée étant que toutes les fluctuations linéaires de X explicables par S ont été captées par le modèle et donc que les résidus ne contiennent plus de dépendance linéaire. De plus, le niveau de suppression de la corrélation peut être contrôlé en introduisant un hyperparamètre α tel que :

$$\tilde{\tilde{X}} = \alpha\tilde{X} + (1-\alpha)X$$

Au-delà de la perte potentielle d'informations que peut engendrer cette méthode, l'absence de corrélation ne garantit pas l'absence de dépendance statistique. La mitigation réalisée est donc limitée. Néanmoins, pour des modèles linéaires tels que les GLM, cette approche reste cohérente pour supprimer toutes les liaisons que peuvent capter les modèles. Il est intéressant de considérer que cette limite puisse aussi représenter une force. En effet, elle permet dans certain cas de mitiger le biais tout en préservant des relations, des informations plus complexes mais essentielles dans le jeu de données.

3.1.3 Adaptation Fair-SMOTE

Le traitement des déséquilibres entre classes est une piste qui peut être exploitée pour la mitigation du biais. Comme discuté dans la section origine du biais, quelle que soit la forme que prend le déséquilibre des classes, il peut conduire à du biais au moment de la modélisation. Une des solutions serait donc de rééquilibrer les classes sur le spectre des distributions concernées.

L'approche classique est de sur-échantillonner la classe sous-représentée de S ou de sous-échantillonner la classe sur-représentée de S. Toutefois, Chakraborty et al.[15] ont prouvé empirement que cette approche amplifie le biais. Cela s'explique par le fait que ces méthodes de rééchantillonnage classique ne tiennent pas compte de la valeur de la variable cible et ne cherche qu'a égaliser la répartition des classes.

Les auteurs proposent de procéder au rééchantillonnage dans chaque sous-groupes constitués de valeurs de Y et de S différentes. Ils appellent cette approche le Fair-SMOTE [1], une méthode construite pour des variables cibles qualitatives. Par exemple, si Y et S sont binaires, il y aura $4 = 2 \times 2$ sous-groupes et un rééchantillonnage est effectué dans chaque sous-groupe. Au final, les classes sont rééquilibrées mais en tenant compte de leur répartition avec la variable cible. Ainsi, au lieu de chercher à faire disparaître les effets de S sur les données ($X \perp\!\!\!\perp S$), l'objectif est plutôt de rééquilibrer $S|Y$. En d'autres termes, pour une valeur quelconque de Y fixée, les individus seraient représentés de la même manière quelle que soit leur valeur de S.

La dernière spécificité de cette approche est la méthode de génération des individus. Au lieu de choisir avec remise les individus initialement contenus dans le sous-groupe, les auteurs rajoutent une subtilité permettant de créer des individus différents dans certain cas. Ils définissent deux hyperparamètres st et ft choisis dans l'intervalle $[0, 1]$, respectivement seuil et facteur de transformation. En partant d'un individu p choisi aléatoirement dans le sous-groupe, un modèle de k plus proche voisin est utilisé pour obtenir les deux individus, v_1 et v_2, les plus proches de l'individu choisi. Les coordonnées du nouvel individu sont reconstruites colonne par colonne de la manière suivante :

— Colonne à valeurs binaires,
  si $st > u$ alors :
    $\tilde{x} = \text{ChoixAleatoire}(x_{v_1}, x_{v_2}, x_p)$
  sinon :
    $\tilde{x} = x_p$.
— colonne à valeurs qualitatives, $\tilde{x} = \text{ChoixAleatoire}(x_{v_1}, x_{v_2}, x_p)$
— colonne à valeurs quantitatives,
  si $st > u$ alors :
    $\tilde{x} = x_p + ft \times (x_{v_1} - x_{v_2})$
  sinon :
    $\tilde{x} = x_p$.

Avec u la réalisation d'une loi uniforme sur $[0, 1]$, $\tilde{x}$ la valeur de la coordonnée du nouvel individu sur la colonne X, $(x_j)_{j \in \{v_1, v_2, p\}}$ la valeur de la coordonnée pour les individus v_1, v_2 et p. Cette méthode permet d'obtenir de nouveaux individus tout en conservant la distribution présente dans le sous-groupe.

Dans le but de pouvoir appliquer cette approche aux cas continus traités dans ces travaux, une adaptation est proposée et testée. La variable d'intérêt Y continue est discrétisée à l'aide des connaissances métiers. Ces connaissances sont appuyées par une étude statistique de la distribution croisée (Y, S). Cela permet de mettre le doigt sur les sous-groupes à rééquilibrer. La méthode fair-SMOTE peut ensuite être appliquée comme décrite dans la littérature. Il est ensuite nécessaire de reconstruire la variable cible pour chaque nouvel individu. Pour se faire, la méthode de création des variables quantitatives est utilisée.

1. SMOTE : Synthetic Minority Oversampling TEchnique, Fair-SMOTE : rééchantillonnage équitable.

A la place des k plus proches voisins, les réseaux antagonistes génératifs (GAN) aurait pu être utilisé soit pour simuler de nouveaux individus dans une approche de type SMOTE, soit pour reconstruire les variables X en générant des individus similaires avec les GAN tout en minimisant la dépendance avec S. Les k plus proches voisins sont préférés pour leur faciliter d'implémentation.

3.2 Mitigation pendant la modélisation

La mitigation pendant la modélisation, précisément pendant l'utilisation des algorithmes de modélisation permet de garder à disposition toutes les informations contenues dans le jeu de données. L'idée étant qu'en conservant toutes ces données, il serait possible d'exploiter le maximum d'informations tout en réduisant les effets provenant de la variable sensible. Ces méthodes sont plus difficiles à mettre en place. En effet, elles demandent un certain accès au processus d'estimation. Il faut ensuite intégrer des contraintes d'équité et enfin assurer la convergence des méthodes.

3.2.1 Exponentiated Gradient

Agarwal et al.[1], proposent une approche de réduction pour la mitigation du biais. Ils montrent d'abord que les définitions d'équité peuvent s'écrire sous forme d'ensembles d'inégalité linéaire de la forme :

$$M\mu(f) \leq c,$$

avec M une matrice, c un vecteur et μ un vecteur de moment conditionnel. Le vecteur c permet de contrôler le niveau auquel chacune des contraintes est mise en place en modifiant les valeurs des c_k. Le programme d'optimisation du problème d'apprentissage statistique devient :

$$\min_{f \in \mathcal{F}} L(h) \text{ sous contrainte que } M\mu(f) \leq c.$$

Les auteurs discutent que l'utilisation de fonctions de prédiction déterministe rend l'application des contraintes plus coûteuses en terme de performance. Ils proposent donc d'utiliser des "fonctions aléatoires". Cela revient à tirer, aléatoirement dans l'ensemble des fonctions définies, une fonction avant chaque prédiction. Ainsi, l'erreur de classification devient $L(Q) = \sum_{f \in \mathcal{F}} Q(f)L(f)$ et les moments conditionnels deviennent $\mu(Q) = \sum_{f \in \mathcal{F}} Q(f)\mu(f)$. Le problème d'optimisation prend ainsi la forme suivante :

$$\min_{Q \in \Delta} L(Q) \text{ sous contrainte que } M\mu(Q) \leq c.$$

Pour résoudre ce problème, ils utilisent l'algorithme de Freund et Schapire[27], Exponentiated Gradient d'où le nom de la méthode. Cette méthode tirée de la théorie des jeux met en compétition la fonction de prédiction et le niveau de respect de la contrainte. L'optimal est atteint lorsque le changement d'un de ces éléments conduit à un gain de performance et d'application de contrainte faible fixé.

Cette méthode est toutefois difficile à généraliser. En effet, pour certaines fonctions de coût et pour toutes les contraintes, il faut remettre en forme le système pour pouvoir l'appliquer à la méthode d'exponentiated gradient. De plus, la fonction de perte doit être 1-Lipschitzienne sous la norme L_1.

3.2.2 Réduction par grid or random search

Dans la continuité de leurs travaux sur la mitigation du biais, Agarwal et al. démontrent que dans le cas où la variable sensible est binaire, il est possible de réduire la recherche de l'optimum à un problème de sélection d'hyperparamètres optimaux. En effet, il est, d'abord, possible de déduire la valeur optimale de la contrainte à appliquer en partant du modèle de prédiction optimal. Ensuite, pour le cas S binaire, il est nécessaire de trouver deux paramètres, un paramètre par contrainte. Ces paramètres étant liés entre eux par une équation, il suffira de trouver la valeur optimale d'un des hyperparamètres pour résoudre l'ensemble du problème.

Ils proposent donc de procéder en utilisant une méthode de recherche par grille pour trouver la valeur optimale. Ainsi, pour un ensemble raisonnable de valeur du paramètre de la contrainte, il trouve les modèles optimaux, et suivant l'arbitrage souhaité entre mitigation et performance, le modèle et les paramètres optimaux sont sélectionnés.

3.2.3 Hyperparamétrage naïf des modèles

Il est possible d'imaginer que certaines configurations d'hyperparamètres et de variables puissent permettre d'obtenir des modèles performants et équitable. L'idée assez naïve sera de procéder à une sélection de variables et/ou à des hyperparamétrages et d'ajouter des métriques de mesure du biais dans les critères de décision.

L'utilisation des réseaux neuronaux est aussi discutée dans la littérature. En effet, ces modèles prédictifs sont modulables et flexibles. Ils permettraient de rajouter des contraintes d'équité à la fonction d'optimisation et de contrôler la convergence. Ils ont toutefois un coût opérationnel non négligeable, les temps de calculs sont longs et les modèles sont difficiles à distribuer en assurance.

3.3 Par ajustement des sorties

La mitigation post modélisation consiste à modifier les prédictions du modèle dans le but de le rendre plus équitable. Ni les données, ni le modèle n'est modifié. Le niveau d'équité après modélisation est mesuré puis un sur-modèle est construit dans le but d'appliquer des règles de décisions plus justes. Le fait de ne pas devoir réentrainer le modèle est un des avantages de cette approche. En effet, réentrainer les modèles peut avoir des coûts en temps, en puissance de calcul et des coûts environnementaux non négligeables pour des modèles de tailles imposantes.
Toutefois, les approches de ce type semblent n'être suffisamment cohérentes que dans le cas Y binaire. Le premier exemple auquel il est facile de penser est celui de la régression

logistique. Les probabilités obtenues permettent de modifier les seuils de décision avec pour conséquence d'impacter le comportement du modèle vis-à-vis des différentes classes. Les méthodes post modélisation présentes dans la littérature sont donc des méthodes de modifications de frontières de décision prenant en compte des mesures d'équité. L'idée est assez intuitive, modifier les règles de décisions pour les classes de variables sensibles dans le but d'améliorer l'équité tout en préservant les performances. Les méthodes suivantes peuvent être citées : equalized odds postprocessing[38], calibrated equalized odds[66] et reject option classification[44].

Il n'existe, au moment de la rédaction de ces travaux, très peu voire aucune recherche accessible et testée sur la mitigation post modélisation pour la régression. Néanmoins, en utilisant le principe de l'équité individuelle, une approche appelée redistribution équitable est proposée. Elle tente de modifier chacune des prédictions dans le but d'obtenir une distribution $\widehat{Y}$ plus équitable.

3.3.1 Redistribution équitable

Dans le cas Y continue, l'adaptation de la méthode Flip-test [2] permet d'obtenir une certaine mesure du biais individuel. Ce biais représente l'écart moyen entre les prédictions d'un individu et les individus les plus proches de la classe opposée. Dans le but de contourner la mise en place de graphe et de modèles causaux, des algorithmes de k plus proches voisins sont utilisés.
Dans le cas du genre par exemple, pour une femme de la base de données, cette méthode permet d'approcher la valeur de Y qu'elle aurait eu si elle était un homme. La qualité de cette mesure est donc fortement dépendante de la capacité de l'algorithme à trouver les bons voisins. Dans le but de réduire au maximum cette contrainte, l'algorithme peut être hyperparamétré et une sélection de variables peut être effectuée dans le but d'obtenir le modèle de k plus proches voisins conduisant aux écarts les plus faibles entre les individus et entre les primes prédites.
Ainsi, à chaque individu i de la base est associé un écart ec_i supposé être le fruit de la variable sensible S. Idéalement, un modèle équitable devrait attribuer à chaque individu la même prédiction que les individus qui lui sont proches sans distinction de genre. Les écarts doivent donc être proches de 0.

La redistribution équitable intervient donc dans le but de redistribuer de manière intelligente les écarts entre les individus de la base de données et ainsi obtenir une meilleure équité. Intuitivement, l'idée serait de corriger directement les prédictions :

$$\widehat{y}_i = \widehat{y}_i - ec_i \text{ ou } \widehat{y}_i = \widehat{y}_i - \frac{ec_i}{2}.$$

La première correction ne fait qu'inverser le signe des écarts entre les différentes classes de S. La seconde ne fonctionne que dans le cas où le nombre de variables explicatives est faible. En effet, dans de grandes dimensionalités, corriger l'intégralité de l'écart dans le but de rapprocher les valeurs de $\widehat{Y}$ peut avoir l'effet inverse et éloigner ces valeurs

2. Voir section 2.5.1

les unes des autres. Aussi, un individu a son voisinage mais appartient également à des voisinages différents suivants les individus qui l'entourent. La correction peut donc rapprocher les valeurs de certains individus en les éloignant significativement des autres. Une approche itérative a prouvé être plus efficace : à chaque itération une part $\frac{ec_i}{\eta}$ de l'écart est réduite en alternant les classes de S. Cet ajustement peut être appliqué pour un nombre d'itérations fixé en avance ou jusqu'à ce que le niveau d'écart global soit en dessous d'un seuil fixé ϵ. Corriger les écarts classe par classe de manière répétée au lieu de les corriger toutes classes confondues permet de s'assurer que les corrections réduisent effectivement les écarts inter-classes respectifs.

Les paramètres η et ϵ sont respectivement la vitesse de correction et la tolérance. Un η petit permet d'avoir une convergence rapide de l'algorithme mais le risque est que l'algorithme fasse de trop grands sauts et diverge dans l'atteinte de la correction optimale. En augmentant η, le temps de calcul devient de plus en plus long. Le choix de ces paramètres est donc un arbitrage entre apport de qualité à la redistribution, mesurée par l'écart global, et le temps de calcul. En plus de vérifier l'indicateur d'écart global, la qualité de la nouvelle distribution de $\widehat{Y}$ doit être examinée. Suivant le cas d'usage, les métriques de performance peuvent suffire. Pour la tarification, les ratios sinistres sur primes (S/P) ainsi que les changements de primes induits doivent aussi être étudiés.

L'algorithme a convergé, les écarts sont réduits et la nouvelle distribution $\widehat{Y}$ est validée, vient maintenant la question de l'utilisation à grande échelle de cette approche. Une grille peut être construite dans le but de donner pour chaque combinaison de modalités les corrections à appliquer. Une autre démarche serait de comprendre à l'aide d'un modèle simple la distribution des écarts définitifs. Par exemple, en utilisant un modèle linéaire pour prédire les écarts, les coefficients obtenus pourraient être utilisés pour ajuster les coefficients GLM.
La première approche soulève la question de la distribution. Il est plus simple de distribuer des coefficients qu'une grande grille contenant tous les ajustements. La seconde approche rajoute un second sur-modèle, ce qui peut être préjudiciable du point de vue du biais statistique. Une solution serait de considérer les deux sur-modèles comme un seul et de valider les résultats directement sur les sorties du modèle linéaire simplifié construit.

La prise en compte de l'équité n'est pas aisé surtout sous contraintes de performance. De plus, tout comme dans le cas des mesures, aucun consensus n'existe sur les approches de mitigation du biais. Le champ de recherches est en pleine expansion et les approches proposées dans la littérature sont dans certains cas trop rattachées au cas d'usage qu'elles traitent. Le cas Y binaire est ici encore le cas de prédilection. Cette section permet tout de même de fournir un spectre relativement complet des approches envisageables et adaptables au cas de la tarification non-vie.
Dans la suite, les différentes approches étudiées sont implémentées dans un cadre d'application le plus réaliste possible. Il s'agit de la tarification d'une garantie d'assurance automobile. Cette tarification est effectuée en prenant en compte les différentes contraintes commerciales et opérationnelles que peuvent imposer l'assurance dans le but d'obtenir une vision concrète de la mise en place de l'équité.

Chapitre 4

Application à la tarification automobile

Cette partie est dédiée à la mise en place de l'équité dans un contexte d'assurance automobile. Ainsi, une tarification est construite en procédant par étape : étude de la problématique, construction de la base, traitements et analyses de données et modélisation de la fréquence et de la sévérité. La modélisation est ensuite enrichie par la construction de zoniers et l'utilisation de véhiculiers. Les modèles sont constamment évalués, améliorés et validés pour permettre d'aboutir aux meilleurs modèles possibles. Une fois cette tarification achevée, la présence du biais est étudiée et des méthodes de mitigation sont mises en place. Cette mitigation a pour but de réduire le biais tout en préservant les performances des modèles construits.
Pour des raisons de confidentialité, les données ont été anonymisées et certains éléments ne seront pas présentés de manière détaillée. Toutefois, cette configuration ne freine en aucune mesure la compréhension et la présentation des résultats sur la thématique de ce mémoire à savoir la détection, la mesure et la mitigation de biais en tarification non-vie. Ces contraintes ne concernant que la partie tarification classique, un domaine assez bien connu de la science actuarielle. De plus, la rédaction et les illustrations assurent aux lecteurs la compréhension et l'analyse des résultats obtenus.

4.1 Tarification d'une garantie en assurance automobile

La tarification en assurance non-vie est un processus long pouvant nécessiter des études sur une période allant de 3 à 9 mois. Dans le processus de tarification, il est utile de prendre en compte l'étude ou la conception de l'offre et l'étude des conditions des contrats avant toute modélisation. Une fois la modélisation terminée, de nombreux ajustements sont appliqués sur les primes et les modèles obtenus. Des études sont aussi menées dans le but de mesurer les impacts de la nouvelle tarification sur le portefeuille assuré et les nouvelles souscriptions. En assurance non-vie, la tarification au sens de modélisation statistique des risques se décompose en deux grandes phases :

1. Modélisation initiale de la fréquence et de la sévérité.
2. Enrichissement de la modélisation (zonier, véhiculier, enrichissement des données sur l'assuré, prise en compte d'interaction etc.) et modélisation finale.

Dans les pages qui suivent, les travaux ayant mené à la construction du tarif sont présentés. Cette étape est indispensable pour fournir une référence en termes de performance et d'équité aux travaux des sections suivantes sur la détection et l'implémentation de l'équité. De plus, l'implémentation de l'équité ne nécessite pas de refondre tout le processus de tarification. Ainsi, les travaux de cette partie sont réutilisés dans la suite. La figure 4.1 illustre les différentes étapes de tarification.

FIGURE 4.1 – Les étapes nécessaires à la tarification en assurance non vie.

4.1.1 Modélisation initiale de la fréquence et de la sévérité

La modélisation initiale de la fréquence et de la sévérité est un problème de modélisation statistique résolu en tenant compte des spécificités du métier d'assureur. Comme pour toute étude statistique, elle débute par une prise de connaissances des enjeux auxquels doivent répondre les modèles. Le but étant de tarifer des produits, une analyse des produits concernés est indispensable. S'en suit une longue phase de construction de bases de données, de variables et de traitement de cette base. Une fois la base préparée, le modèle est construit, optimisé en sélectionnant les variables, en choisissant les bonnes lois et les bons hyperparamètres. Il est fréquent d'effectuer plusieurs allers-retours entre la phase de modélisation et la phase de traitement des données.

Analyse des besoins commerciaux. Avant d'entamer toute tarification, l'analyse des besoins commerciaux est indispensable. Dans cette étape, les équipes actuarielles peuvent être amenées à participer à la conception du produit, à l'étude des garanties proposées et de la cible du produit. Cette étape permet de comprendre les besoins aux-

quels doit répondre la modélisation. En certains points, les connaissances acquises permettront de guider les choix dans la construction des variables, les retraitements ou les ajustements de primes. Dans le cas présenté dans ces travaux, l'objectif est de construire une tarification automobile pour la garantie bris de glace, une garantie vendue comme une composante des formules de garanties. Après la garantie obligatoire qui est la responsabilité civile, il s'agit de la garantie la plus commercialisée par l'assureur avec une présence dans 87% des formules vendues.

La tarification la plus performante possible est recherchée suivant certaines contraintes qui sont précisées dans la présentation des retraitements et des choix de modélisation. Une fois cette tarification construite, la mesure et la mitigation des biais sont implémentées.

Prise en main et pré-traitement des bases de données. Les produits nécessitant une tarification étant cernés, les bases de données nécessaires à cette tâche doivent être construites. Pour cela, il faut éventuellement consolider des données provenant de plusieurs systèmes d'informations, choisir l'historique et l'étendue en cohérence avec l'étude. La figure 4.2 présente l'ensemble des variables à disposition en les regroupant suivant 4 axes : l'assuré, son bien, son contrat et les informations sur les sinistres.

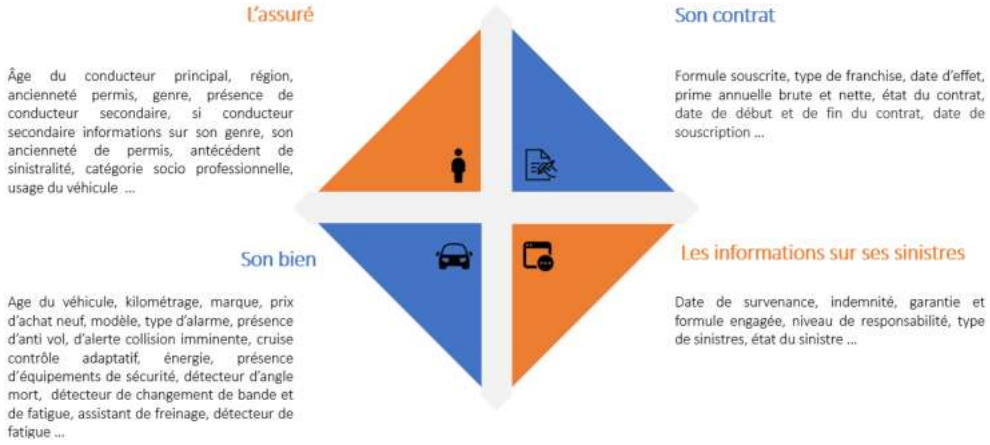


FIGURE 4.2 – Les variables à disposition regroupées suivants 4 axes.

Deux bases principales sont utilisées : une base de contrats et une base de sinistres. La base de contrats contient toutes les informations sur le contrat, son souscripteur et le bien couvert. La base de sinistres, quant à elle, contient une liste de tous les sinistres. Ces bases regroupent des données qui s'étendent de 2016 à 2021, soit 5 années d'historique. Sur cette période 221000 individus ont été en portefeuille pour un total de 189000 années d'exposition.

Initialement, la base contrat contient 55 variables. Les traitements suivants sont appliqués :

— Prise en compte des formats dates dans la base de données. Une attention particulière est accordée à ces dates pour assurer leur cohérence. Par exemple, la date de fin de contrat se chevauchant avec la date de début de contrat.
— Calcul de l'ancienneté de la police et du véhicule, et de l'exposition des polices. Ces calculs sont effectués pour chaque année d'historique.
— Prise en compte des revalorisations tarifaires.
— Prise en compte des formules de garanties. Pour chaque police, il a fallu calculer une exposition par rapport aux garanties souscrites. A l'aide d'une table de correspondance qui permet d'avoir la liste des garanties qui composent une formule, les garanties souscrites ont pu être isolées et une exposition a été calculée en partant des résultats des étapes précédentes.

Une base donnant les caractéristiques des véhicules (un équivalent à la base SRA) est utilisée pour enrichir les données contrats. Il s'agit de données permettant d'avoir un maximum d'informations sur le véhicule assuré. Il offre des compléments d'information tels que le type d'énergie, le nombre de places, le poids du véhicule, le nombre de cylindres, le type de boîtes etc. Le rapprochement entre la base de contrat et la base décrivant véhicules est effectué en utilisant le numéro de série du véhicule. Pour cette jointure, 18% des polices de la base de données n'ont pas eu de correspondance dans la base des véhicules. Pour ces polices restantes, un rapprochement par modèle du véhicule est utilisé. Après cette seconde jointure, il ne reste que 0,6% des polices sans correspondance. Ces polices sont rapprochées en utilisant les modèles de voitures les plus proches dans le véhiculier, le tout en minimisant les écarts de prix.
Pour ces polices, l'écart de prix moyen était de 3400€ et 95% des polices ont des écarts de prix de moins de 6000€. La figure 4.3 résume la jointure de la base contrat et la base décrivant les véhicules.

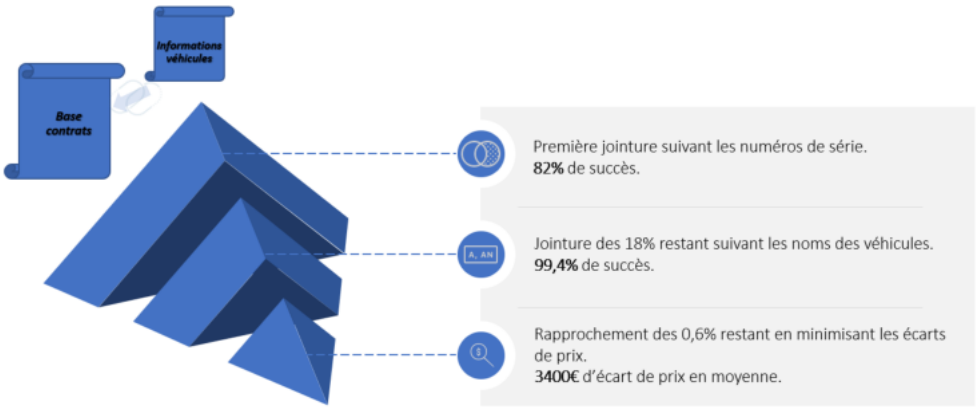


FIGURE 4.3 – Illustration de la jointure entre base contrat et informations véhicules.

La base sinistre contient initialement 32 variables. Les traitements suivants sont appliqués :
— Reformatage des variables, mapping des modalités et traitement des doublons.

— Calcul des années de survenance, de la charge totale et du coût moyen.

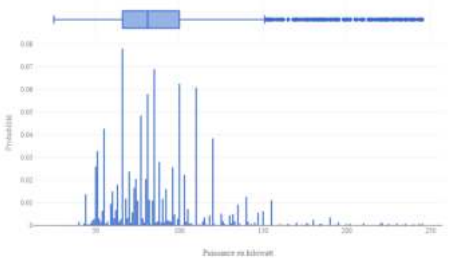
Une première pré-sélection des variables permet de ne retenir que celles qui sont cohérentes avec les objectifs de tarification. Le tableau 4.1 décrit brièvement ces variables. Ce tableau contient 4 colonnes : nom de la variable, description, valeur prise et statistique. La colonne statistique présente la moyenne et la médiane pour les variables quantitatives. Pour les variables qualitatives, les parts des deux modalités les plus représentées sont présentées.

Nom de la variable	Description	Valeur prise	Statistique
charge	charges de sinistres individuelles	$[0, +\infty[$	$\bar{x}$: 620€ | Me : 645€
nbr	nombre de sinistre	$[0, 4[$	0 : 96% | 1 : 3,3%
expo	exposition par contrat	[0,27%, 100%[	$\bar{x}$: 78% | Me : 96%
annee_pol	année de souscription de la police	$[2016, 2021[$	2019 : 19% | 2018 : 17%
age_cp	âge du conducteur principal	$[18, 90[$	$\bar{x}$: 51 | Me : 50
anc_cp	nombre d'année de présence en portefeuille	$[0, 13[$	$\bar{x}$: 1 | Me : 1,78
region	code région	21 zones	C : 34% | A : 27%
genre	genre du conducteur principal	F, M	M : 60,4%
anc_permis_cp	nombre d'année d'ancienneté du permis	$[0, 54[$	$\bar{x}$: 30 | Me : 29
conducteur_2	présence de conducteur secondaire	0, 1	0 : 77%
age_veh	âge en année du véhicule	$[0, 62[$	$\bar{x}$: 5 | Me : 4
energie	type d'énergie	5 types	D : 53% | E : 45%
poids_kg	poids en kilogramme	$[865, 3551[$	$\bar{x}$: 1380 | Me : 1355
kw	puissance du véhicule en KW	$[13, 245[$	$\bar{x}$: 89 | Me : 82
klm_souscrit	forfait kilométrage	$[12k, 50k[$	$12k$: 43% | $18k$: 15%
prix_cata	prix en catalogue	$[6800, 65[$	$\bar{x}$: 22900€ | Me : 19425€
type_boite	type de boîte de vitesses	2 types	A : 93%
ante_sinis	antécédent de sinistre	0, 1	0 : 90,8%

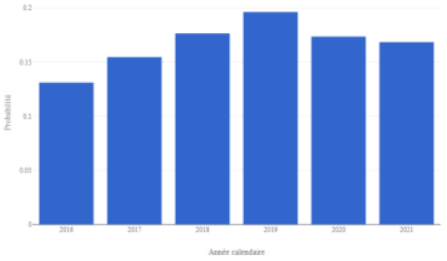
TABLE 4.1 – Description des variables du jeu de données.

Une fois ces bases importées et pré-traitées, pour permettre leur utilisation, il est nécessaire de les analyser et de peaufiner les traitements.

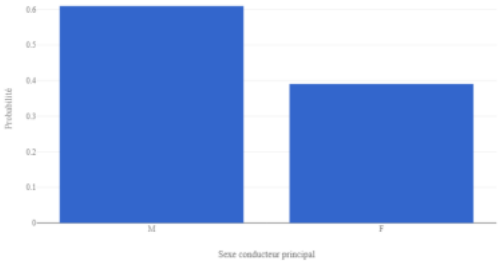
Analyse et traitement des données. Une analyse approfondie est effectuée sur les bases obtenues. Cette analyse débute par une étude descriptive des variables, l'étude des distributions et des répartitions dans les classes. Une attention particulière est accordée à la détection d'anomalies et de valeurs extrêmes. Il faut aussi mesurer l'influence des variables prises individuellement sur les variables d'intérêts (fréquence et coût moyen des sinistres). Il s'agit d'une analyse univariée. Les tests d'hypothèses sont des outils pouvant servir à mesurer la significativité des observations empiriques. Ensuite, une analyse multivariée est menée dans le but d'étudier la corrélation entre l'ensemble des variables, en utilisant le V de cramer, le test d'indépendance de χ^2 ou les méthodes factorielles (ACP, ACM). La figure 4.4 met en avant quelques illustrations des travaux d'analyse de données.



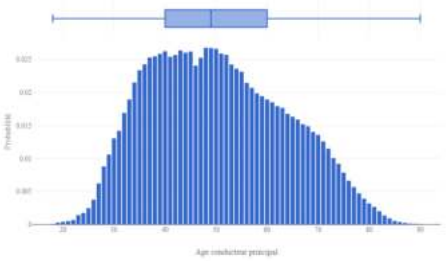
(a) Distribution de la puissance du véhicule



(b) Répartition du nombre de contrat en cours par année



(c) Répartition du nombre de contrat par genre



(d) Distribution de l'âge du conducteur principal

FIGURE 4.4 – Quelques illustrations d'analyse de données.

Après avoir pris connaissance des données, il est nécessaire de procéder à des traitements supplémentaires. Dans cette étape, les anomalies présentes dans le jeu de données sont corrigées. Les valeurs manquantes sont traitées en faisant soit de l'imputation soit de la suppression de données. Le traitement des valeurs manquantes est toujours sujet à débat ; l'imputation pouvant introduire du biais dans la base de données et la suppression pouvant conduire à une perte d'informations. Les valeurs extrêmes sont repérées et des décisions doivent être prises quant à leur intégration dans la modélisation. En effet, en tarification non-vie, les sinistres peuvent avoir des distributions à queue lourde. Une distinction entre sinistres graves et sinistres attritionnels est importante car leur présence dans la base de données peut totalement changer le comportement du modèle. La théorie des valeurs extrêmes ou les quantiles peuvent permettre la définition du seuil à partir duquel les sinistres sont considérés comme "graves". Il est aussi possible de fixer ce seuil à dire d'expert ou suivant la structure de réassurance choisie. Une fois les extrêmes repérés, un modèle spécifique pourra ensuite être construit pour cette sous base. Un modèle généralement plus simple du fait de la faible volumétrie de données. Il est aussi possible de répartir les charges de sinistres graves sur l'ensemble des primes du portefeuille. Quelle que soit la méthode choisie, la prime finale sera une somme entre la prime attritionnelle et la prime grave.
Dans certains cas, des franchises ou montants de remboursement forfaitaires peuvent être présents. Ces éléments conduisent à une présence importante de certains montants dans

la base des coûts de sinistres.

Dans la base de sinistres étudiée, les montants forfaitaires ne représentent que 2% des coûts totaux et 5,1% des sinistres. Ces montants sont donc retirés de la base. Les franchises et valeurs extrêmes représentent 0,8% de la base pour un total de charge de 1,2%. Ces sinistres sont écrêtés de la modélisation. La figure 4.5 montre la distribution des charges de sinistres pour la garantie bris de glace.

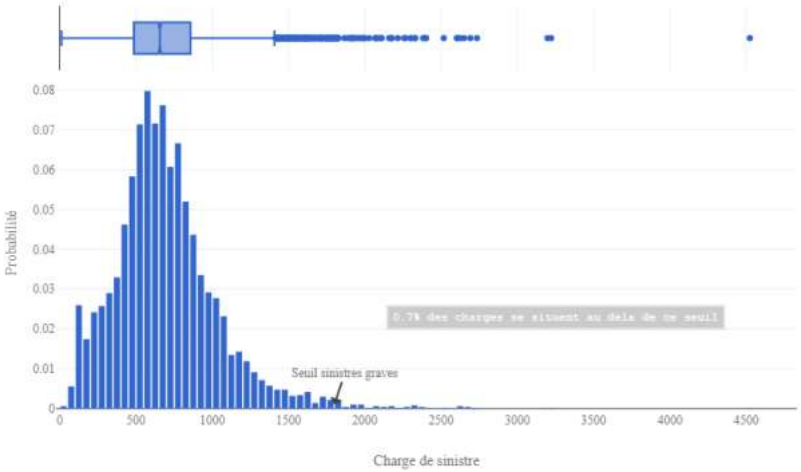


FIGURE 4.5 – Distribution des charges pour la garantie bris de glace.

Au final, 5,1% des sinistres sont retirés pour un coût total de 3,2%, extrêmes, franchises et forfaits compris. L'approche adoptée est de ne pas prendre en compte ces éléments dans la modélisation pour ensuite les rajouter comme une surprime sur tous les individus. Le retraitement final prendra la forme suivante :

$$\text{prime pure finale } i = \text{prime pure prédite } i + \frac{\text{charges totales exclues}}{\text{expo totale}}.$$

Dans la suite, cet ajustement sera appliqué au moment de mesurer l'équilibre tarifaire obtenu avec un modèle. En effet, la somme des primes prédites est comparée à la somme des sinistres historiques.

En ce qui concerne les valeurs manquantes, elles sont traitées par imputation. Ainsi, pour les variables qualitatives, les modalités les plus représentées sont imputées. Pour les variables quantitatives, une imputation par la moyenne a été initialement utilisée. Une fois ces variables discrétisées, la modalité la plus représentée est imputée. Ces imputations ont introduit un biais statistique négligeable dans les modélisations car, pour les variables finales retenues, la part de valeurs manquantes n'était pas significative ($\leq$ 1%, voire nulle).

Après ces traitements, une étude des variables d'intérêts par année donne les résultats affichés dans le tableau 4.2.

annee_pol	cout_moyen	freq	prime_pure	charge_total	expo_total	nbr_pol
2016	504,3	0,0755	31,5	445331,53	15899,5	20027
2017	523,1	0,0765	35,6	648622,86	18366,4	23172
2018	552,9	0,0647	32,4	648692,74	24213,1	30623
2019	530,4	0,0578	32,0	772326,09	26345,6	31563
2020	570,8	0,0412	24,4	641561,31	23883,1	29741
2021	613,4	0,0424	25,7	433920,34	21134,0	28530

TABLE 4.2 – Statistiques sur les variables d'intérêts.

Il apparaît des effets observables sur l'ensemble du marché automobile : une hausse des coûts moyens principalement due à une augmentation des coûts des pièces détachées et une baisse de la sinistralité en 2020 et 2021 due à la COVID 19. Les statistiques publiées en 2022 par la Fédération Française de l'Assurance mettent en avant une hausse de 15% des coûts moyens entre 2019 et 2021.

Dans les modélisations, les années ne sont pas prises en compte. Les données sont retraitées dans le but d'obtenir une cohérence entre les statistiques et les objectifs de l'assureur. En règle générale, plusieurs approches sont envisageables.
Pour le traitement de la hausse des coûts moyens, le modèle de fréquence est maintenu mais les coûts sont ajustés à l'aide d'indices permettant de quantifier la hausse générale des coûts de réparation. Ainsi, tous les coûts sont retraités pour avoir les valeurs qu'ils auraient eu au moment de la construction du tarif. Cette méthode est appelée le "as-if".
Pour le traitement de l'effet covid, le modèle de coût est maintenu mais pour le modèle de fréquence, les années 2020 et 2021 peuvent être supprimées de la base de modélisation. Une autre approche est d'utiliser toutes les années dans le modèle et ensuite ne retenir que les coefficients d'une année de référence. Après retraitements, les statistiques agrégées sont les suivantes :

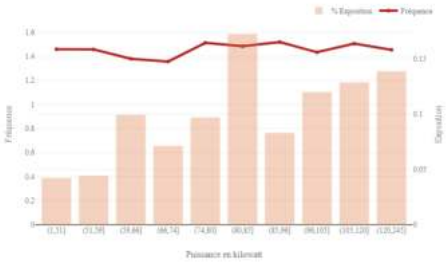
coût moyen : 573,18€, fréquence moyenne : 5,165% et prime moyenne : 27,661€.

Construction des variables de modélisations. Avant la construction des modèles, il est nécessaire de discrétiser les variables continues et de regrouper les modalités des variables discrètes. Ces discrétisations permettent de construire des classes de risques homogènes, de lisser les sauts de primes entre des classes assez proches et d'introduire un caractère non linéaire dans certains modèles utilisés plus bas. Le but étant de construire une segmentation cohérente avec les risques du portefeuille mais aussi commercialement acceptable. L'actuaire peut utiliser des méthodes comme le K-means ou la Classification Ascendante Hiérarchique (CAH) dans le but de construire ces classes tout en utilisant ses connaissances d'expert. Ainsi, en partant des connaissances métiers et d'une analyse de la sinistralité et de l'exposition, des classes sont construites. De la sensibilité des variables d'intérêt aux fluctuations de la variable explicative, la cohérence des découpages ou regroupements et l'amélioration de la performance des modèles sont les facteurs les plus pris en compte pour la construction des variables. Tous ces facteurs sont maximisés sous contraintes de la parcimonie du modèle et donc du nombre de modalités. Il est important

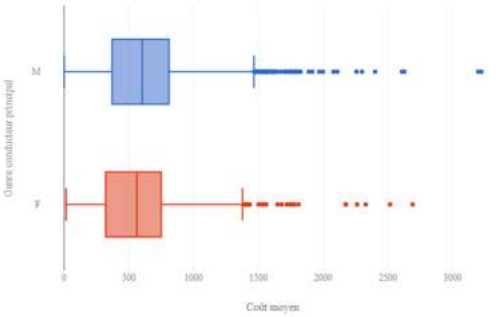
de noter que de nombreux allers-retours ont lieu entre les phases de modélisation et de traitement dans le but d'obtenir le meilleur traitement possible.

La figure 4.6 illustre la sensibilité des variables d'intérêt à certaines des variables explicatives. Des matrices de corrélations sont construites à plusieurs étapes des traitements pour comprendre les liens existants entre les variables et vérifier que les différents traitements ne détruisent pas ces liens.

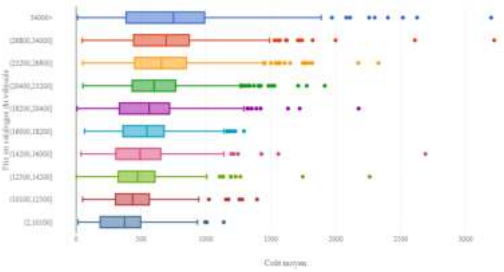
(a) Fréquence et exposition par ancienneté du permis



(b) Fréquence et exposition par puissance du véhicule



(c) Coût moyen par genre



(d) Coût moyen par prix en catalogue du véhicule

FIGURE 4.6 – Quelques illustrations d'analyse de données sur les variables définitives.

Modélisation de la fréquence et de la sévérité. Ces retraitements ont permis de construire une base de données. Il est maintenant question de modéliser la prime pure. Cette modélisation repose en règle générale sur le modèle collectif. Ce modèle stipule que la somme des sinistres du portefeuille est une somme aléatoire de variables aléatoires indépendantes et identiquement distribuées représentant chacune un coût de sinistres. Ainsi, ce modèle nous donne la prime pure comme étant une distribution composée fréquence-sévérité :

$$S_i = \sum_{j=1}^{N_i} C_{i,j}$$

avec N_i la variable aléatoire représentant le nombre de sinistres de l'assuré i, $C_{i,j}$ le coût du $j^{\text{ème}}$ sinistre. Les variables $Y_{i,j}$ sont indépendantes et identiquement distribuées $\forall i, j$.

Dans la pratique, l'assureur modélise séparément la fréquence de sinistres par assuré et le coût unitaire de chaque sinistre. Il va donc constituer deux sous bases, chacune permettant la modélisation d'une variable d'intérêt. Les variables cibles étant le nombre ou la charge de sinistres, le cadre de modélisation est celui de la régression. Depuis 1990, la classe des modèles utilisée est celle des GLM (Generalized Linear Models). Ces modèles permettent d'étendre le cadre de la régression linéaire multiple en rajoutant une fonction de lien et des lois différentes. En notant :

- g la fonction de lien ;
- Y_i la variable cible ;
- X_i la matrice des variables explicatives, où chaque colonne est une variable explicative ;
- β le vecteur des paramètres à estimer.

Le modèle s'écrit :

$$g(\mathbb{E}[Y_i|X_i]) = X_i^t\beta$$

Ce modèle est un modèle pseudo linéaire dans la mesure où il permet de prendre en compte certaines relations non linéaires à l'aide de la fonction de lien et de la distribution choisie. Ces modèles ont fait l'objet d'études mathématiques extensives et complètes. Ils sont simples à utiliser et interprétables (sous réserve de la parcimonie des modèles). Ils permettent en quelques manipulations de construire une grille tarifaire et ne nécessite donc pas une exécution du modèle pour chaque prédiction.

Comme évoqué plus haut, ces modèles permettent de spécifier une loi et une fonction de lien. Les lois utilisées sont des lois de la famille exponentielle. Pour le modèle de fréquence, il est nécessaire d'utiliser une loi de comptage. En effet, le nombre de sinistre est une variable discrète positive à valeurs entières. Les lois utilisables sont la loi de Poisson et la loi Binomiale Négative. Le coût des sinistres est une variable à support positif. Les lois utilisables sont donc la loi Gamma et la loi Lognormale. Il est aussi possible de modéliser directement la charge totale de sinistres en utilisant des distributions composées Poisson-Gamma issues de la sous-famille des lois exponentielles : les Tweedies. Ces lois ont la particularité d'avoir une masse non nulle en 0 tout en restant continues sur le reste de la distribution.
Les GLM présentent toutefois des limites telles que la non prise en compte native des effets croisés entre les variables, la linéarité de la relation entre variables explicatives et variable cible. Ils souffrent des défauts de leurs qualités : la simplicité ! Ils sont malgré tout les modèles qui s'imposent sur le marché et ne semblent pas être en passe d'être remplacés. En effet, ces dernières années, l'utilisation de modèles plus complexes a pris de l'ampleur. Toutefois, leur interprétabilité et les difficultés de mise en production limitent leur utilisation.

Dans ces travaux, en plus du modèle GLM, les algorithmes boîtes noires de forêts aléatoires et de gradient boosting sont utilisés. Ils ont prouvé empiriquement leurs performances sur un large spectre de problèmes statistiques et sur des travaux assurantiels. Dans la suite, les algorithmes de forêt aléatoire et de gradient boosting sont notés respectivement RF et GB. En ce qui concerne les GLM, la loi la Gamma est utilisée pour

le coût moyen, la loi de Poisson pour les fréquences et la loi Tweedie pour les primes pures. Pour ces trois modèles la fonction de lien *log* est utilisée. Ces hyperparamètres ont prouvés être les plus performants sur ce cas d'usage. Il est important de préciser qu'il n'existe pas de critère universel permettant de définir la notion de meilleur modèle. Les meilleurs modèles sont donc déterminés par rapport aux critères qui sont définis pour les comparer. Les trois critères utilisés sont les suivants :

1. $$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i(\widehat{y}_i - y_i)^2}.$$

2. $$MAE = \frac{1}{N}\sum_{i=1}^{N} w_i|\widehat{y}_i - y_i|.$$

3. $$True/pred = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i \widehat{y}_i}.$$

Avec w_i le poids associé à la variable d'intérêt. w_i est l'exposition pour la fréquence et la prime pure, et le nombre de sinistre pour le coût moyen. N est le nombre d'individus de la base de test. En effet, pour permettre une évaluation pertinente des modèles construits, une subdivision aléatoire et stratifiée des bases de données initiales est implémentée. Cette approche permet de s'assurer que les distributions sont quasiment les mêmes entre base de test et base d'apprentissage. Ces évaluations sont menées dans une méthode de validation croisée à cinq segments. Une subdivision temporelle aurait pu être envisagée. Toutefois, cette approche aurait pu conduire à des différences de distributions. En plus de cela, les variables d'intérêts ont été retraitées pour ne plus tenir compte des années. Ainsi, le choix est fait d'utiliser une approche aléatoire stratifiée.
Les résultats des différents modèles sur les différentes variables d'intérêts sont rapportés dans les tableaux 4.3, 4.4 et 4.5.

Évaluation, amélioration et validation du modèle. Après la construction des modèles avec le choix des lois et l'estimation des paramètres, la dernière étape consiste à évaluer, améliorer et valider le modèle. Des métriques comme le AIC, RMSE, GINI, R^2 ajusté et la déviance sont utilisées pour évaluer la précision des modèles et les comparer entre eux. Des procédures de sélection de variables sont utilisées dans le but d'obtenir des modèles plus parcimonieux et précis. Une fois les modèles les plus satisfaisants sélectionnés, il est indispensable de les valider. D'abord, du point de vue statistique, en étudiant les résidus et en vérifiant que les hypothèses sous-jacentes aux modélisations sont cohérentes. Puis, du point de vue métier, en étudiant la cohérence des coefficients obtenus et de l'ensemble des primes.

Cette étape conduit à des traitements de données et à la sélection de variables. Partant des variables pré-sélectionnées présentées dans le tableau 4.1, les variables kw, poids_kg

Modèles	RMSE	MAE	True/Pred
GLM	323,94	247,15	1,061
GB	320,16	245,03	1,028
RF	321,25	246,77	1,033

Table 4.3 – Performances des modèles de coût.

Modèles	RMSE	MAE	True/Pred
GLM	0,301	0,137	0,957
GB	0,299	0,134	0,961
RF	0,294	0,131	0,962

Table 4.4 – Performances des modèles de fréquence.

Modèles	RMSE	MAE	True/Pred
GLM	179,49	57,96	1,034
GB	183,01	60,23	1,051
RF	179,34	57,55	1,038

Table 4.5 – Performances des modèles de prime pure.

et age_cp sont retirées. En effet, la variable age_cp est fortement corrélée avec les variables anc_permis_cp et age_veh. Le choix est fait de retirer la variable age_cp.
Les variables kw et poids_kg sont remplacées par une variable poids_kw ; un ratio entre poids et puissance. Ce remplacement permet de retirer deux variables significativement corrélées tout en conservant la pertinence des informations qu'elles contiennent.

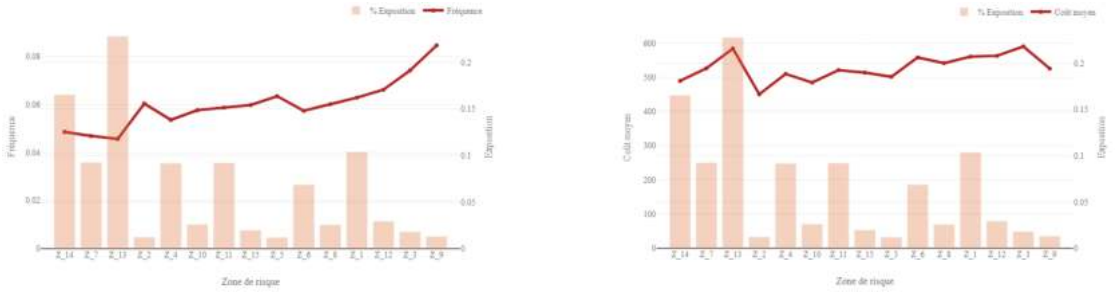
Les modèles sont aussi hyperparamétrés. Cet hyperparamétrage est effectué en deux étapes. Après une sélection avisée des champs de valeurs possibles, un premier hyperparamétrage est effectué de manière aléatoire en utilisant les distributions de valeurs possibles pour les hyperparamètres. Cette approche, moins coûteuse en temps de calcul, permet de trouver plus rapidement les sous-espaces d'hyperparamètres optimaux. Une recherche exhaustive est ensuite effectuée sur les périmètres restreints obtenus.

4.1.2 Enrichissement de la tarification

Cette phase pourrait être intégrée dans la partie modélisation mais cela ne ferait pas justice au travail spécifique requis. En règle générale, les travaux de cette phase conduisent à la construction de variables qui sont intégrées dans le modèle de la première phase.

Construction de zoniers. Dans l'ensemble des données à la disposition de l'assureur se trouvent des informations géographiques : l'adresse de l'assuré et les données externes (open data) sur les zones géographiques. Un zonier est une variable qualitative dont chaque modalité représente une zone de risque. Cette variable permet de résumer l'information relative aux comportements/fluctuations géographiques du risque.

Un zonier peut être construit pour le modèle de fréquence des sinistres, ou pour celui des coûts des sinistres ou pour ces deux modèles à la fois. Une fois les modèles initiaux de fréquence et de sévérité construits avec des données ne contenant aucune information géographique, les erreurs de prédictions de ces modèles doivent être récupérées. Il s'agit de la différence entre les prédictions et les valeurs réelles. Des transformations peuvent être



(a) Fréquences et expositions par zone de risque

(b) Coûts moyens et expositions par zone de risque

FIGURE 4.7 – Quelques statistiques par zone de risque.

appliquées sur ces résidus. Sous l'hypothèse que la variable d'intérêt dépend de facteurs géographiques, l'absence de ces facteurs dans les modèles devrait conduire à la présence d'informations géographiques dans les résidus des modèles. L'assureur peut soit expliquer ces résidus à l'aide de variables externes ou les lisser dans le but de former des zones.

Il est donc indispensable de s'assurer que les résidus peuvent être expliqués partiellement par des variables géographiques. Une fois cette hypothèse vérifiée, le choix de la maille de modélisation s'impose. La maille peut être départementale, communale, à l'adresse etc. Un lissage IDW (Inverse Distance Weighted) ou par krigeage est appliqué dans le but de combler les zones du maillage ne contenant pas ou peu de données. Les résidus lissés sont ensuite regroupés en des zones en utilisant des quantiles, des k-means ou des CAH.

Pour la détermination du nombre de zones, un arbitrage est effectué entre précision de la segmentation et besoins commerciaux. En effet, plus il y a de zones, plus le découpage est précis mais plus les clients peuvent se sentir lésés par ces fluctuations de primes. Les méthodes de sélection optimale du nombre de zones favorisent l'utilisation d'un grand nombre de zones tandis que l'assureur peut pour des raisons commerciales vouloir réduire au maximum ce nombre. Ainsi, suivant les garanties et les objectifs de l'assureur, par exemple privilégier l'équilibre tarifaire ou les assurés, le nombre de zones retenues est généralement différent. Dans cette démarche, une prise en compte de la crédibilité des résidus est importante. En effet, plus une composante de la maille est exposée au risque plus les observations y sont de qualité et vice versa.

Un zonier unique est construit en partant des résidus des modèles de coût et de fréquence obtenus par le premier modèle hors zone géographique. Ce zonier permet de définir 15 zones de risque. La méthode k-means est utilisée. Un nombre de zones élevé est retenu car la finesse du découpage est l'élément le plus important pour le portefeuille étudié. Ainsi, certaines zones ayant des expositions relativement faibles, sont tout de même retenues. La figure 4.7 présente les coûts moyens, les fréquences et les expositions par zone de risque.

Prise en compte des interactions. Les GLM présentent l'inconvénient de ne pas prendre en compte par défaut les effets croisés entre les variables. Le seul moyen de les

prendre en compte est de les intégrer à la main dans les modèles. Une solution serait de calculer toutes les interactions possibles entre les variables et de faire une sélection de variables à l'aide de méthodes automatiques. Néanmoins, cette solution a un coût algorithmique exponentiel. Il y a donc un réel enjeu à étudier les interactions et à sélectionner les plus pertinentes qui permettront l'amélioration de la modélisation. Une des approches envisageables est de construire des modèles prenant en compte les interactions telle que la forêt aléatoire, de détecter les interactions les plus pertinentes à l'aide de méthodes d'interprétabilité et d'analyse de sensibilité et d'intégrer ces interactions dans les modèles de bases.

L'intégration d'interactions n'est pas assez significative dans le cas étudié, le ratio coûts sur bénéfices n'est pas favorable à leur intégration. Il faut toutefois reconnaître que des études plus poussées auraient pu être menées sur ce sujet. Des indices de Sobol ou des effets croisés par valeurs shapley auraient peut être permis de détecter des interactions plus significatives.

Le cas particulier du véhiculier. Le véhiculier pourrait se construire de manière similaire au zonier, en utilisant les résidus des modèles sans informations sur les véhicules. Toutefois, pour notre étude, ce véhiculier n'a pas été construit. Certaines variables caractérisant le type de véhicule couvert sont retenues directement dès la première modélisation.

Les tableaux 4.6, 4.7 et 4.8 donnent les métriques définitives à l'issue de la phase de construction du tarif. Ces résultats sont donc les références de performances atteignables avec le jeu de données à disposition en tenant compte des contraintes métiers.

Modèles	RMSE	MAE	True/Pred
GLM	306,72	239,63	1,016
GB	304,21	238,81	1,019
RF	304,27	238,81	1,022

TABLE 4.6 – Performances des modèles de coût

Modèles	RMSE	MAE	True/Pred
GLM	0,271	0,112	0,970
GB	0,272	0,109	0,970
RF	0,267	0,112	0,971

TABLE 4.7 – Performances des modèles de fréquence

Modèles	RMSE	MAE	True/Pred
GLM	171,08	52,94	1,010
GB	172,33	53,07	1,023
RF	171,04	52,94	1,017

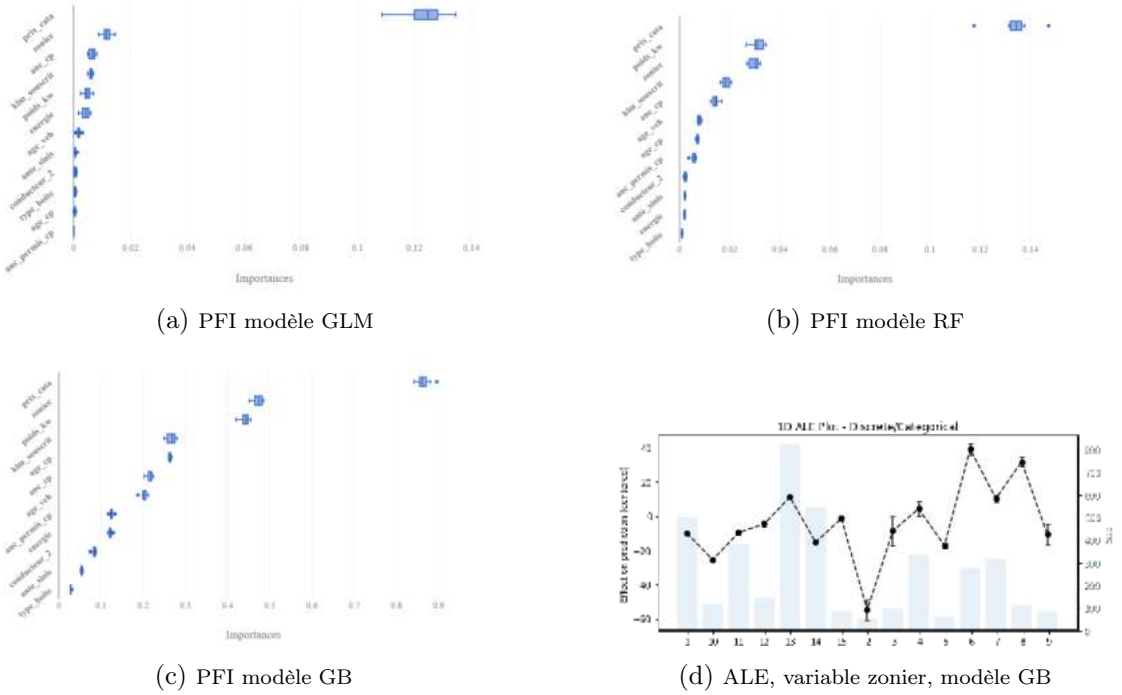
TABLE 4.8 – Performances des modèles de prime pure

4.1.3 Interprétabilité des modèles implémentés

L'interprétation des modèles construits est une part importante de la modélisation. Elle permet de comprendre comment fonctionne le modèle pour faire ses prédictions,

de déceler des problèmes ou découvrir des éléments qui n'étaient pas visibles dans la phase d'analyse. Les GLM étant nativement interprétables, leurs coefficients peuvent être utilisés pour l'interprétabilité. En ce qui concerne les modèles dit boîtes noires tels que RF et GB, les Accumulated Local Effects, Permutation Feature Importance (resp. ALE et PFI) et valeurs shapley sont utilisés pour les interpréter. Le lecteur pourra se référer à l'annexe A pour avoir plus de détails sur ces méthodes.

Analyse globale des modèles. Cette analyse permet de comprendre les variables clés utilisées par les modèles. Pour ce faire, les méthodes PFI et ALE sont utilisées. La méthode PFI calcule un indice d'importance des variables pour la prédiction. Elle évalue à quel point l'information présente dans la variable est pertinente pour obtenir les prédictions. La méthode ALE évalue l'effet cumulé d'une variable sur la prédiction moyenne. La figure 4.8 présente quelques résultats de l'étude globale des modèles de coût.



(a) PFI modèle GLM

(b) PFI modèle RF

(c) PFI modèle GB

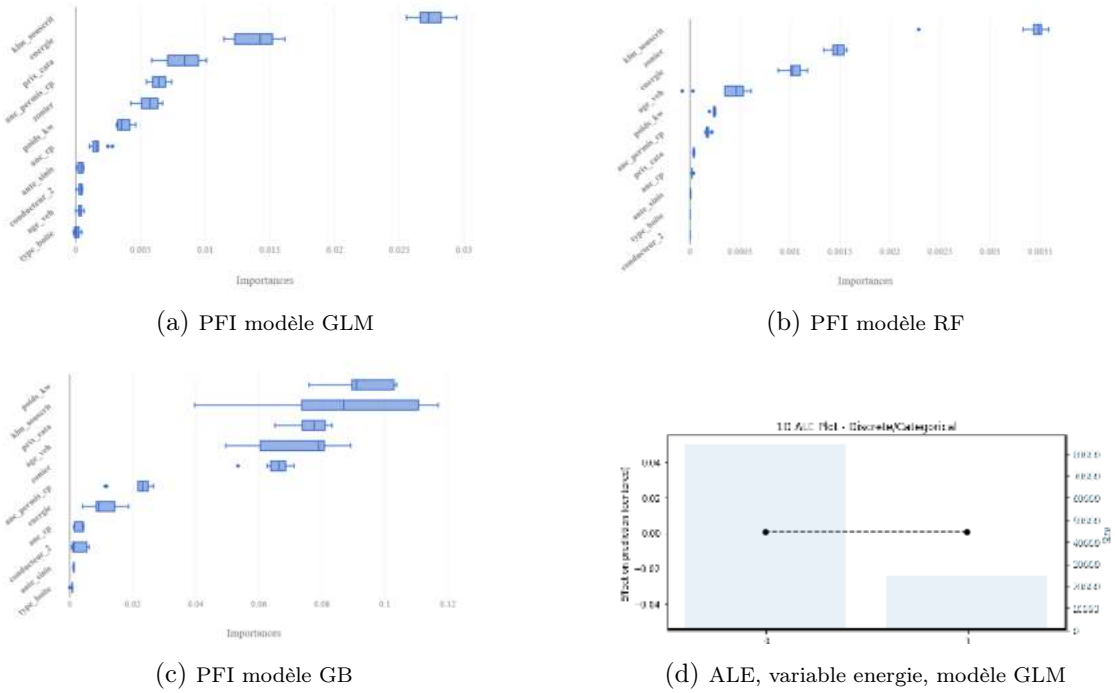
(d) ALE, variable zonier, modèle GB

FIGURE 4.8 – Modèle de coût : effets et importances des variables avec PFI et ALE.

L'analyse des sorties des modèles et des différents graphiques permet de remarquer que le prix en catalogue du véhicule est sans contexte la variable la plus significative pour ces différents modèles de coût. Des variables comme zonier, poids_kw et anc_cp contribuent ensuite de manière plus ou moins significative suivant les modèles. Les variables type_boite, ante_sinistre, conducteur_2, energie et anc_permis_cp sont des variables moins significatives. Ces observations sont assez cohérentes. En effet, sachant qu'un sinistre a eu lieu, son coût moyen est la somme du coût du matériel et de sa pose. Il est

intuitif de penser que plus le véhicule est coûteux plus ses composantes le sont aussi. Le graphique ALE s'interprète de la manière suivante : l'effet moyen de l'appartenance à la zone 6 sur les prédictions du coût moyen est de +40 et l'effet moyen de l'appartenance à la zone 1 sur les prédictions du coût moyen est de -10.

La figure 4.9 présente quelques résultats de l'étude globale des modèles de fréquence. Pour les modèles de fréquence, le kilométrage souscrit apparaît comme étant la variable



(a) PFI modèle GLM

(b) PFI modèle RF

(c) PFI modèle GB

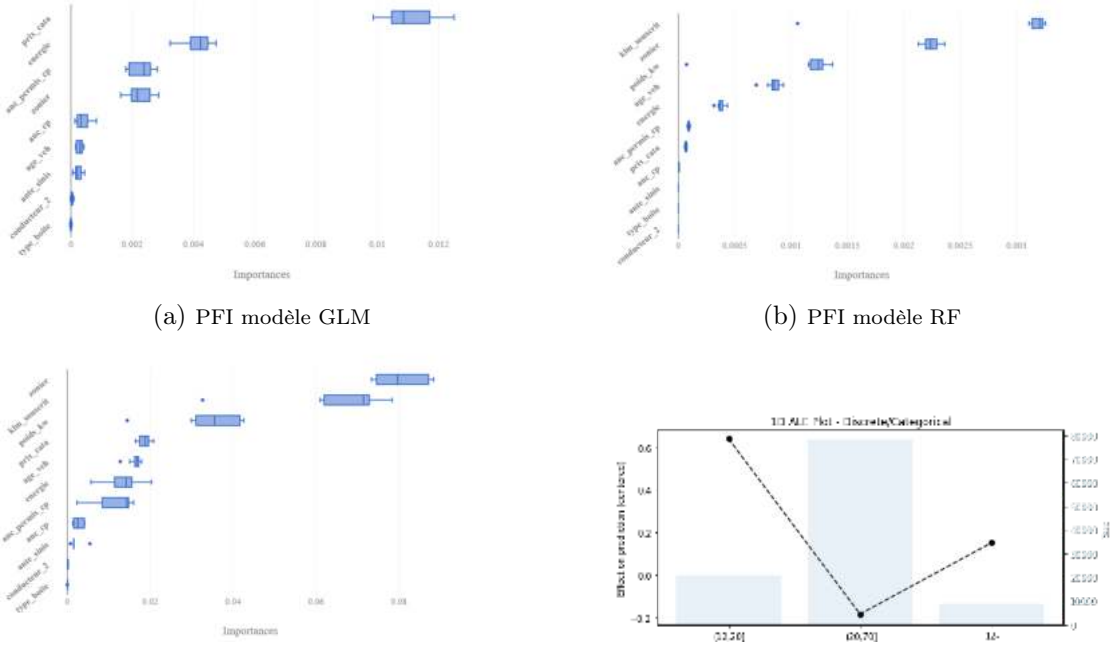
(d) ALE, variable energie, modèle GLM

FIGURE 4.9 – Modèles de fréquence : effets et importance des variables avec PFI et ALE.

la plus significative. Il s'agit d'une observation cohérente : plus l'assuré utilise son véhicule, plus il a de chances d'avoir un sinistre et plus il souscrit à un kilométrage élevé. Ces deux variables sont donc des co-facteurs. Ceci est une nouvelle illustration intéressante des relations de causalité qui peuvent exister dans les données. En effet, augmenter le kilométrage souscrit ne conduit pas directement à une augmentation de la fréquence de sinistres. Il serait trompeur de dire que ces variables sont directement corrélées. En réalité, l'augmentation du kilométrage est liée aux habitudes de conduite (nombre de kilomètres parcourus en moyenne) tout comme la fréquence de sinistres. Ainsi à travers la variable non observée nombre de kilomètres parcourus en moyenne, par exemple, les variables kilométrage et fréquence sont liées.

La figure 4.10 présente quelques résultats de l'étude globale des modèles de prime pure. Attention, pour des raisons de mise à l'échelle, les importances de variables klm_souscrit

et poids_kw ont été retirées du graphique PFI GLM. Elles se classaient respectivement en première et seconde position en terme d'importance PFI.
Pour la prédiction des primes pures, les modèles semblent utiliser des variables différentes. Il apparaît toutefois que les variables zonier, klm_souscrit, poids_kw et prix_cata semblent être les variables les plus significatives : un mixte des variables importantes pour les modèles de coût et de fréquence.



(a) PFI modèle GLM

(b) PFI modèle RF

(c) PFI modèle GB

(d) ALE, variable anc_permis_cp, modèle RF

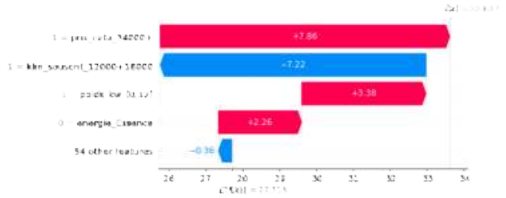
FIGURE 4.10 – Modèles de prime pure : effets et importances des variables avec PFI et ALE.

Analyse locale des modèles. Cette analyse permet de comprendre les prédictions prises individuellement ainsi que les contributions de chaque variables à la prédiction obtenue. Pour cela, les valeurs de Shapley sont utilisées. Ces valeurs ont été introduites en théorie des jeux par Shapley en 1953[72].
Étant donné un jeu coopératif où p joueurs collaborent dans le but d'obtenir un gain, survient alors le problème de l'attribution équitable du gain aux p joueurs du jeu. En d'autres termes quel est l'apport de chaque joueur au gain obtenu ? C'est dans le but de répondre à cette problématique que les valeurs de Shapley ont été introduites. Le parallèle est fait entre le cadre théorique initial de la théorie des jeux et celui de l'interprétabilité des modèles. Le jeu est la prédiction d'une certaine variable cible à l'aide d'un vecteur de valeurs (une instance d'un jeu de données), les joueurs sont les valeurs des variables explicatives et le gain est la différence entre la prédiction et la valeur moyenne

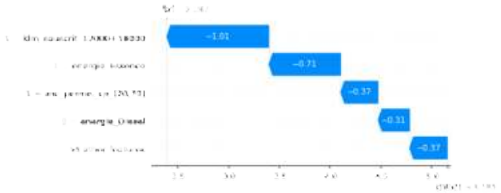
de la variable cible dans le jeu de données.
Ainsi, la valeur de Shapley de la variable j dans l'obtention du gain g se définit comme étant sa contribution (au gain total) additionnée et pondérée sur toutes les combinaisons possibles des valeurs des variables explicatives.

La figure 4.10 présente les résultats des valeurs de Shapley pour certaines prédictions.

(a) Décomposition additive de la prédiction du coût



(b) Décomposition additive de la prédiction de la prime pure



(c) Décomposition additive de la prédiction de la fréquence (en %)

FIGURE 4.11 – Décompositions additives à l'aide des valeurs shapley.

Pour la prédiction individuelle de la prime pure, en partant de la prime moyenne de 27,715€, la prédiction de 33,637€ est obtenue en additionnant les effets des variables prix_cata, klm_souscrit, poids_kw et energie. L'écart entre la valeur prédite et la valeur espérée s'explique par l'effet de ces 4 variables. Les autres variables jouent un rôle minime dans le niveau de la prédiction (−0,36€).
En règle générale, les variables les plus importantes détectées par les méthodes globales apparaissent comme étant les plus importantes pour déterminer le niveau des prédictions. Néanmoins, d'un individu à un autre, le rôle des variables dans la prédiction peut totalement différer. Il est important de garder à l'esprit que les résultats de ces méthodes sont des résultats moyens.
L'exemple affiché pour une prédiction du coût moyen montre que le prix_cata est la variable qui contribue le plus à la déviation de la valeur moyenne. Cette observation est cohérente avec les importances des variables globales. Derrière cette variable, les variables anc_cp et age_cp sont celles qui contribuent le plus alors qu'elles n'apparaissaient pas comme étant les plus significatives. Cela s'explique par le fait que l'individu vit dans une zone de risque moyen et a un ratio poids_kw moyen. Ainsi, ces variables apparaissent car les variables les plus discriminantes en moyenne ne le sont pas pour cet individu.
L'interprétation est la même pour la prédiction de la fréquence. L'écart entre la moyenne de 5% et la prédiction de 2% s'explique comme la somme des contributions données par

le graphique. Du fait des fondements théoriques solides qui soutiennent cette méthode, la décomposition additive ainsi obtenue peut être acceptée par les régulateurs et utilisée pour décrire le niveau de prime. Un des grands inconvénients de cette méthode est le temps de calcul qu'elle requiert.

Les modèles sont construits et optimisés. Dans une construction classique de tarif, il ne resterait plus qu'à éventuellement faire les derniers retraitements des paramètres. L'intérêt de ces travaux est d'aller plus loin et de proposer une étude des effets de variables dites indésirables, ce qui a été appelé biais. Ces biais sont étudiés, mesurés et mitigés dans une démarche permettant la reproductibilité pour d'autres variables sensibles.

4.2 Mesure du biais

Le cadre choisi pour cette étude du biais est celui de la variable genre. C'est le cas d'usage de variables sensibles le plus connu en science actuarielle. Les variables d'intérêt sont le coût moyen, la fréquence et la prime pure.
L'objectif est d'étudier l'effet du genre sur les différentes variables d'intérêts avant et après modélisation. Étant donné qu'aucune discrimination sur le genre ne doit être admise dans les primes, ses effets sont jugés indésirables d'où le terme de biais précédemment introduit. Ce biais fait référence à un biais éthique. Dans cette section, il est étudié dans les moindres détails.

4.2.1 Mesures du biais sur les données historiques

Dans le but de quantifier l'effet de la variable sensible sur les variables d'intérêts, des mesures de dépendance sont utilisées. En effet, une indépendance entre Y et S signifierait que quasiment aucun biais n'existe dans les données. Le seul biais qui pourrait être retranscrit sur les prédictions serait le biais lié aux décisions de modélisation.
Pour mesurer ces effets, six mesures de dépendance sont choisies. En effet, choisir plus de mesures conduirait à une utilisation inefficace du temps à disposition pour interprétation et arbitrage entre les mesures. Par définition, suivant les cas d'usages, certaines métriques sont plus appropriées. Par exemple, quand la variable d'intérêt est binaire, les métriques p-rule, impact disparate ou maltraitance disparate sont utilisées dans la littérature. Le binaire en assurance peut apparaître lorsque le contrat étudié prend fin après le premier sinistre ou quand l'assureur doit classer ses risques en bons ou mauvais risques ou encore détecter des cas de fraude. Les cas continus restent toutefois les plus présents. De plus, ces cas sont plus difficiles à aborder. En effet, les matrices de confusion croisées avec la variable sensible permettent aisément de prendre en main le cas de Y binaire.

Pour le cas continu ici présenté, les métriques utilisées sont méticuleusement choisies pour fournir une vision globale des différents aspects de la dépendance entre les variables. Les raisons ayant motivé les choix de chaque métrique sont les suivantes :

— **Tau de Kendall :** il correspond aux métriques classiques de mesures de dépendance. Il est assez connu et malgré ses limites, il permet d'avoir une première idée

du niveau de dépendance présent.

— **Ratio des moyennes :** il permet d'avoir une première idée compréhensible des écarts qui peuvent exister entre la distribution des hommes et celle des femmes. La moyenne est toutefois sensible aux valeurs extrêmes.

— **p-value du test de Kolmogorov-Smirnov :** ce test est l'un des tests usuels de comparaison de distributions de probabilité. Il fournit, comme les deux premières métriques, une base relativement bien connue. Ce test a toutefois un désavantage, il est sensible aux déviations extrêmes. En effet, de par sa définition, il ne prend en compte que le plus grand écart de probabilité. Ce test sera donc une manière de quantifier l'évolution des dépendances dans les extrêmes.

— **La divergence de JS :** cette métrique permet de mesurer la distance entre deux distributions en se basant sur toute l'étendue des distributions. En présence de quantités suffisantes de données, les variations les plus extrêmes sont noyées. Cette distance fournie donc un complément intéressant au résultat du test de Kolmogorov-Smirnov.

— **HGR KDE :** cette métrique est la plus solide théoriquement. Elle permet de mesurer tout type de dépendances entre tout type de variables. Deux versions ont été implémentées ; la version basée sur les réseaux neuronaux et la version utilisant les noyaux (KDE). La version KDE a été retenue car elle converge deux fois plus rapidement que celle basée sur les réseaux neuronaux. De plus, le choix de la structure optimale du réseau de neurones est délicat. Une structure plus légère permettrait une convergence plus rapide mais éventuellement des résultats moins fiables.

— **Adaptation flip-test :** cette métrique permet de mesurer l'équité individuelle sans nécessiter les constructions d'un graphe et d'un modèle causal. Cela est possible en contrepartie de l'hypothèse qu'une distance ou qu'un algorithme permette de mesurer la proximité entre les individus de la base de données. Initialement construite pour une variable cible binaire, elle a été adaptée dans ces travaux en utilisant un algorithme de k plus proches voisins comme mesure de proximité. Ainsi, il est possible de comparer la prédiction d'un individu à celles des individus du genre opposé les plus proches. La qualité de la notion de proximité fournie par le modèle de k plus proches voisins construit est la principale limite de cette approche. Pour contrôler cette limite, un hyperparamétrage et une sélection de variables sont menés. La distance choisie et son paramétrage ainsi que le nombre de voisins sont optimisés. Le nombre de variables et les variables retenues pour construire le modèle sont aussi optimisés. L'objectif étant d'obtenir le modèle de k plus proches voisins conduisant à la plus petite distance entre les individus et au plus faible écart de primes. Ainsi, sur ces différents paramètres, une recherche par grille est effectuée et les résultats sont vérifiés sur une base de test. Bien qu'optimisé, il se pourrait que ce modèle soit moins performant qu'un modèle causal. Cette approche fournit toutefois une solution satisfaisante et ayant un coût opérationnel significativement plus faible que celui qu'engendrerait une étude causale.

Le tau de kendall et le HGR KDE sont des mesures de la force de dépendance. Ils donnent

respectivement des valeurs entre $[-1, 1]$ et entre $[0, 1]$ qui permettent de savoir à quel point deux variables sont dépendantes. Les variables sont donc "indépendantes" suivant ces mesures si elles sont proches de 0. Le HGR étant positif, il ne permet toutefois pas de connaître le sens de la dépendance.

Le test K-S, la divergence de JS et le ratio des moyennes comparent les distributions de $Y|S$. Dans le cas étudié où S est le genre, la distribution des coûts moyens des hommes est comparée à celle des femmes. Il est clair que si $Y \perp\!\!\!\perp S$, alors :

$$\mathcal{L}(Y|\{S : \mathrm{F}\}) = \mathcal{L}(Y|\{S : \mathrm{H}\}) = \mathcal{L}(Y).$$

Ainsi, plus la distance entre ces distributions est grande plus la dépendance entre les variables est significative. Pour qu'il y ait indépendance, il faut donc que le test de K-S soit accepté, que la divergence de JS soit proche de 0 et que le ratio de moyenne soit proche de 1.

Le flip-test évalue la différence entre les distributions en intégrant la notion de proximité entre les individus. Pour qu'il y ait indépendance, il faudrait que cette mesure soit proche de 0. Cela signifierait que les hommes et les femmes du même voisinage ont les mêmes valeurs de Y.
Les cinq premières mesures sont des mesures d'équité de groupe, plus précisément d'indépendance. En effet, les différences de primes entre homme et femme étant connues et présentes dans l'historique, la définition du biais doit se faire en faisant abstraction de Y. Ainsi, contrairement à la suffisance et à la séparation, l'indépendance est la seule définition ne tenant pas compte de Y dans sa caractérisation du biais.
La dernière mesure est une mesure individuelle. Elle tente de mesurer le biais que subit chaque assuré pris individuellement, cela permet de voir le biais sous un angle différent.
Le tableau 4.9 présente les résultats pour les mesures de dépendances entre chacune des variables d'intérêts et la variable sensible.

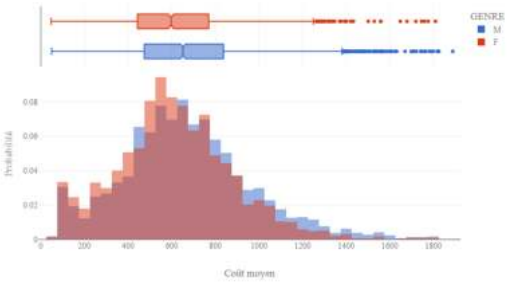
Y	Kendall	HGR_KDE	K-S	Div_JS	Ratio_moy	Flip-test
Coût moyen	-0,0594	0,0816	1,8721e-07	0,3317	1,1023	-12,41€
Fréquence	-0,0182	0,3163	3,3376e-02	0,8198	1,2389	-2,14%
Prime	-0,0183	0,3168	2,9522e-02	0,8188	1,3350	-10,78€

TABLE 4.9 – Dépendance entre Y et S avant modélisation.

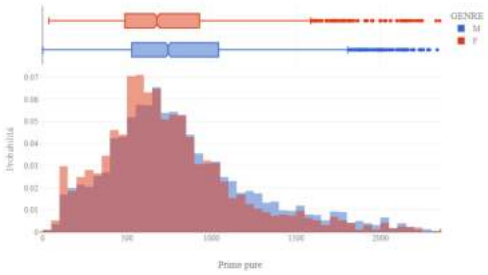
Le tau de Kendall détecte une très faible dépendance entre les Y et les S. Le genre ayant été recodé 0 pour masculin et 1 pour féminin, le signe négatif signifie que les variables sensibles sont plus faibles pour les femmes. HGR KDE détecte déjà des dépendances un peu plus fortes ce qui est cohérent car par définition il capte un plus grand spectre de dépendances. Le test de K-S est rejeté et il semble exister des différences non négligeables entre les distributions homme et femme. Le flip-test nous permet de dire que par rapport aux assurés masculins similaires, une femme a un coût moyen plus faible de 12€. Il s'agit

d'une observation cohérente avec le signe du tau de Kendall. La figure 4.12 permet une analyse des distributions de ces variables d'intérêt par rapport au genre.

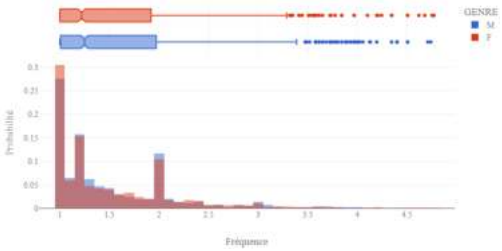
Remarques : Pour permettre la visualisation des fréquences, celles comprises entre [0,1[sont retirées car elles écrasent le reste de la distribution. Sur cet intervalle, les répartitions hommes-femmes sont les mêmes avec, à peu près, 95% de la population distribuée de manière similaire. En 1, il y a 3,4% de femmes contre 2,9% d'hommes.
La notion de prime pure présenté sur l'historique représente la distribution "coût moyen × fréquence". Conceptuellement, ce n'est pas une prime pure car elle ne tient compte d'aucune mutualisation. Une solution aurait été d'approcher la prime historique en calculant des primes moyennes sur des segments de la population constitués par les croisement entre plusieurs variables. Cette approche introduit toutefois du biais car les primes obtenus dépendent des segments choisis. Le choix est fait d'ignorer cette solution et de garder la distribution initiale. Ce choix est cohérent avec l'idée de mesurer les biais sur les distributions se trouvant en entrée des modèles. Sur le graphique, les valeurs nulles sont retirées.



(a) Coût moyen en fonction du genre



(b) Prime pure en fonction du genre



(c) Fréquence en fonction du genre

FIGURE 4.12 – Distribution des Y en fonction du genre.

Il apparaît que les distributions des femmes sont un peu plus asymétriques à droite que celles des hommes. Ainsi, dans les données historiques, en coût moyen, en fréquence et en prime pure, les hommes ont des valeurs plus élevées en moyenne. En observant les graphiques, il peut sembler que les différences de distributions sont plus marquées pour les coûts moyens alors que les mesures ne conduisent pas aux mêmes conclusions. En

effet, sur le graphique des coûts moyens, il y a plus de zones où les distributions ne se chevauchent pas. Alors que la divergence de JS est de 0,33 pour le coût et de 0,88 pour la fréquence.
Il suffit de regarder les échelles de probabilité pour comprendre ces différences. Puisque les écarts sur les distributions de fréquences et de primes sont associés à des écarts de probabilité plus grands, leurs pondérations sont donc plus grandes et conduisent à des écarts plus significatifs dans les mesures.

Il reste difficile d'interpréter dans l'absolu les chiffres obtenus à l'aide de ces mesures. En effet, du fait du manque d'antécédent dans ce type de problématique, il est impossible d'établir des seuils critiques. Néanmoins, ces premiers résultats laissent apparaître les signes de l'effet du genre sur les différentes variables d'intérêts. Ainsi, en plus du déséquilibre des classes (plus d'hommes que de femmes), il existe un biais dans les données historiques. Cette différence est bien connue en tarification automobile et est traitée en omettant le genre ou en rééquilibrant les sorties des modèles.

4.2.2 Biais après modélisation

Tous les résultats présentés par la suite sont obtenus à l'aide des modèles GLM. Bien que les modèles boîtes noires puissent obtenir de meilleures performances, les GLM restent les plus répandus car ils sont simples à interpréter et à intégrer dans les outils de tarification en production. Ainsi, la légère hausse de performances obtenue à l'aide du modèle RF ne suffit pas pour le sélectionner du fait des coûts opérationnels que cela engendrerait. Toutefois, à chaque étape, les calculs sont effectués sur les trois modèles. En général, les tendances sont les mêmes quel que soit le modèle, les écarts de valeurs ne sont pas significatifs.

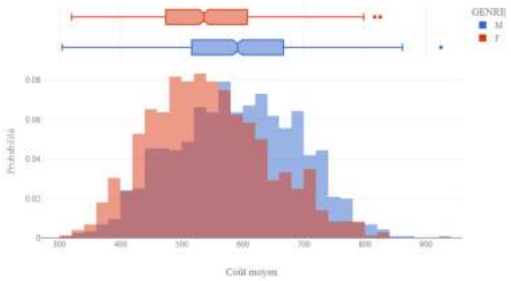
Modélisation avec le genre dans le modèle. Par définition, le modèle contenant le genre serait le plus performant car il exploite toutes les informations à la disposition de l'assureur pour la modélisation du risque. Toutefois, c'est aussi le modèle le plus inéquitable par rapport au genre car une distinction entre homme et femme est directement perceptible dans les primes obtenues. Ce modèle est construit en réinjectant le genre dans les modèles précédemment étudiés. Une fois construites, les prédictions $\widehat{Y}$ sont récupérées et la dépendance entre S et $\widehat{Y}$ est mesurée. Le tableau 4.10 regroupe les résultats de ces différentes mesures.

$\widehat{Y}$	Kendall	HGR_KDE	K-S	Div_JS	Ratio_moy	Flip-test
Coût moyen	-0,1820	0,2137	9,1994e-18	0,3914	1,0861	-3,16€
Fréquence	-0,1849	0,3197	0,0000e+00	0,7493	1,2573	-1,12%
Prime	-0,2018	0,3378	0,0000e+00	0,7226	1,3524	-1,21€

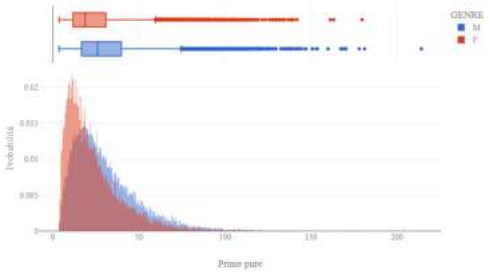
TABLE 4.10 – Dépendance entre $\widehat{Y}$ et S après modélisation contenant le genre.

Les dépendances entre les variables d'intérêt et la variable sensible se sont amplifiées suivant quasiment toutes les mesures. Les mesures exploitant les distributions telles que

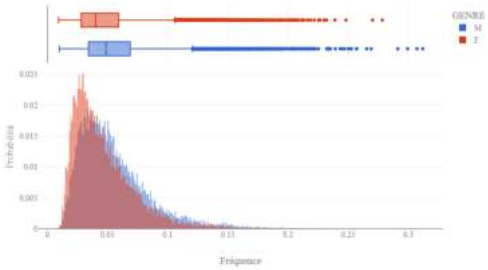
div_JS et le flip-test donnent des dépendances plus faibles car $\widehat{Y}$ est moins dispersée que Y. Le tau de Kendall capte trois fois plus de dépendance pour le coût. Cela voudrait dire que la structure de dépendance est un peu plus simple à détecter. Des tests ont permis de vérifier que cette amplification de la dépendance était la même quel que soit le modèle choisi. Les modèles construits n'ont donc pas uniquement répliqué le biais historiquement présent dans les données, ils l'ont aussi aggravé. La figure 4.13 permet une analyse des distributions de ces variables d'intérêt prédites par rapport au genre.



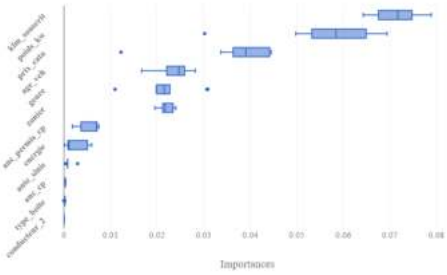
(a) Coût moyen en fonction du genre



(b) Prime pure en fonction du genre



(c) Fréquence en fonction du genre



(d) PFI, modèle GLM

FIGURE 4.13 – Distributions des $\widehat{Y}$ en fonction du genre.

Le décalage entre les distributions s'est amplifié après modélisation. Il y a une plus grande partie des distributions qui ne se chevauchent pas. Les différences apparaissent plus clairement contrairement au cas des données historiques.

Comment expliquer cette amplification du biais ?
L'interprétation des modèles construits permet de remarquer que le genre occupe une place importante dans les processus de prédiction. La sous figure 4.13d en est une illustration.

Bien que le genre soit une variable importante, pourquoi le biais mesuré est-il supérieur à celui présent dans l'historique utilisé pour la modélisation ?
Cela s'explique par l'interdépendance entre les variables du jeu de données. En effet, il existe un lien entre les variables explicatives retenues et les variables d'intérêt ; c'est parce

qu'elles permettent de comprendre le risque qu'elles ont été retenues. Toutefois, en plus de leurs capacités prédictives, ces variables peuvent être liées entre elles. Par exemple, la puissance du véhicule est plus ou moins liée à son prix.
Ces interdépendances font partie des éléments les plus malicieux de la mise en place de l'équité car les variables dites sensibles sont en général des variables exerçant une influence non négligeable sur les distributions des autres variables observées. Ainsi, même si à première vue, les interdépendances observables semblent faibles, leurs accumulations conduisent à accroître le rôle de la variable sensible dans les modèles. Dans les données utilisés, le genre affecte de manière non négligeable la distribution de la puissance, du poids, du type de boîte de vitesse, de la formule souscrite et du prix du véhicule, sans compter ses relations mineures avec les autres variables explicatives. La figure 4.14 permet de visualiser le fait que le genre exerce une influence directe sur les variables d'intérêt mais aussi un effet indirect non négligeable provenant de ses liens avec les autres variables. Ainsi, dans le modèle de coût par exemple, le genre est la 5ème variable la

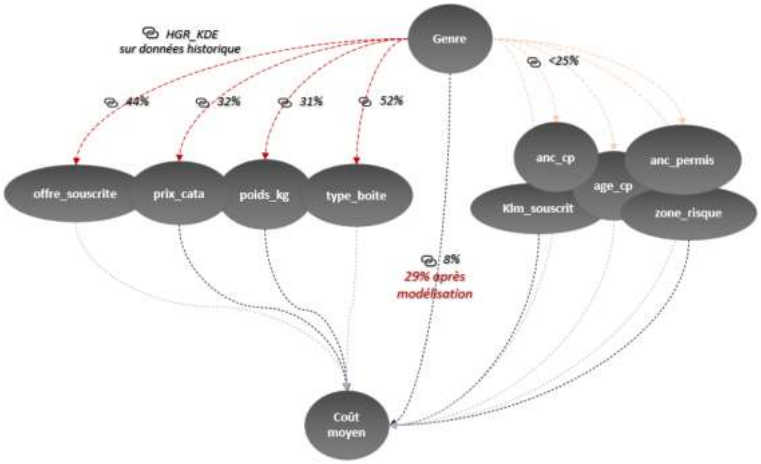


FIGURE 4.14 – Les effets directs et indirects du genre sur les variables du jeu de données.

plus importante d'après la méthode PFI. Toutefois, les variables les plus significatives sont des variables liées au genre. De manière générale, les variables sensibles affectent le comportement des individus. Par exemple, l'âge a une influence sur la prise de risque, le niveau de responsabilité, les centres d'intérêts etc. et l'ethnie peut être étroitement liée au lieu d'habitation, aux services souscrits, aux préférences culturelles etc.

Ces diverses analyses conduisent inéluctablement à un unique questionnement : que se passe-t-il une fois que le genre est omis ou retraité dans la modélisation ? En effet, peut-être que ces interdépendances sont juste sur-interprétées et que les modèles se sont trop appuyés sur le genre.

Modélisation sans le genre dans les modèles. Par définition, le modèle ne contenant pas le genre est un modèle qui permet d'éviter la discrimination directe : les résultats

obtenus in fine ne seront pas différenciés par genre. Il a été discuté en introduction du chapitre sur les mesures de biais que cette approche ne permet pas de traiter la discrimination indirecte car elle ne tient pas compte des liens entre les autres variables explicatives et le genre.
Le tableau 4.11 affiche les résultats pour la dépendance entre $\widehat{Y}$ et le genre pour les modèles ne contenant pas le genre comme variable explicative.

$\widehat{Y}$	Kendall	HGR_KDE	K-S	Div_JS	Ratio_moy	Flip-test
Coût moyen	-0,173418	0,2041	8,8360e-17	0,3888	1,0814	-2,89€
Fréquence	-0,132525	0,2801	5,7498e-237	0,6624	1,1718	-1,03%
Prime	-0,174112	0,2971	0,0000e+00	0,7055	1,2892	-0,90€

TABLE 4.11 – Dépendance entre $\widehat{Y}$ et S après modélisation ne contenant pas le genre.

Les résultats obtenus sont sensiblement les mêmes que ceux obtenus lorsque le genre est présent dans le modèle, il y a légèrement moins de biais. Ainsi, malgré son absence des modèles, le genre influence toujours autant les différentes modélisations.
A première vue ces résultats peuvent sembler non intuitifs : les primes ne sont plus distinguées par genre mais il existe toujours une distinction entre les primes par le genre. Toutefois, en repensant à l'exemple simplifié de la section 2.2, les interdépendances indirectes permettent de faire des distinctions entre les genres. Par exemple, en appliquant des primes différentes pour la présence de conducteur secondaire, celles-ci ayant des répartitions différentes par genre, des primes différentes sont indirectement appliquées au genre.
Dans des modèles multivariés comme ceux construits ici, ces distinctions entre genres peuvent provenir de combinaisons entre plusieurs variables. Par exemple, une combinaison puissance de véhicule, type de boîte et zone de risque peut être très déséquilibrée en termes de répartition homme-femme.
A titre d'illustration, en croisant zonier et prix des véhicules, des déséquilibres significatifs apparaissent entre les genres dans le jeu de données. 87% des individus ayant un véhicule coûtant plus de 30000€ et vivant en zone 12 sont des hommes. Ainsi, toute différence significative de prime entre ce segment et les autres impacte fortement les hommes.

Une approche de modélisation différente. Ces résultats montrent à quel point l'implémentation de l'équité par omission ne résout pas le problème de la discrimination. Dans le but de vérifier ces niveaux de dépendances, une approche de modélisation différente est implémentée. Au lieu de retirer la variable genre du modèle, celui-ci est construit avec elle puis les résultats sont retraités en sortie pour assurer une tarification insensible au genre. C'est l'une des approches adoptée par les assureurs pour faire face à la gender directive. L'approche retenue pour ce retraitement est de calculer une moyenne pondérée des sorties hommes et femmes pour tous les segments contenus dans les modèles. Par exemple, en supposant que le modèle de fréquence est construit avec 2 variables, le genre et une variable F contenant deux modalités F1 et F2, le tableau

4.12 illustre le retraitement effectué en sortie du modèle. La fréquence corrigée de 5,06% s'obtient en effectuant le ratio suivant : $\frac{5,6*450+4,4*363}{450+363}$.

Genre/F	F1	F2	
F	5,60%	3,70%	*fréquence*
	450	562	*somme exposition*
H	4,40%	4,70%	*fréquence*
	363	477	*somme exposition*
Fréquence retraitée	5,06%	4,15%	

TABLE 4.12 – Illustration du principe de retraitement des fréquences après modélisation.

Une fois les retraitements effectués, la dépendance entre ces nouvelles valeurs de $\widehat{Y}$ et S est mesurée. Le tableau 4.13 contient les résultats de ces différentes mesures.

$\widehat{Y}$	Kendall	HGR_KDE	K-S	Div_JS	Ratio_moy	Flip-test
Coût moyen	-0,1810	0,2098	4,1707e-18	0,3862	1,0803	-2,33€
Fréquence	-0,1308	0,2751	7,269e-221	0,6519	1,1723	-0,93%
Prime	-0,1724	0,3063	0,0000e+00	0,7111	1,2525	-0,81€

TABLE 4.13 – Dépendance entre $\widehat{Y}$ et S après retraitement du genre à la sortie des modèles contenant le genre.

Les résultats sont sensiblement les mêmes. Cette approche semble toutefois réduire un peu plus les différences entre les individus semblables ; le biais détecté par le flip-test est légèrement plus faible.
Au global, cette approche ne permet pas de résoudre le problème induit par les interdépendances car en repondérant les sorties, les déséquilibres existants entre les différents segments sont maintenus.
Ainsi, l'effet du genre perdure dans la modélisation même après qu'il soit omis ou retraité de manière classique. Dans la littérature, certains auteurs parlent de la reconstruction de la variable sensible et donc du maintien du biais. Il y a reconstruction lorsqu'en présence d'interdépendance entre les variables explicatives et la variable sensible, ces variables explicatives permettent de reproduire les effets de la variable sensible dans les modèles et donc de perpétrer le biais. Cela semble être le cas dans cette étude.

Reconstruction du genre. Pour vérifier empiriquement l'idée suivant laquelle les autres variables explicatives permettent de reconstruire le genre, des modèles prédictifs sont construits. L'objectif est de tenter de prédire le genre des individus à l'aide des autres variables explicatives. La précision des modèles peut servir comme proxy de la force de reconstruction de S. Pour cette expérience, la base contenant tous les assurés est utilisée. Les variables explicatives utilisées sont celles retenues pour les modélisations. Leur forme définitive est utilisée, c'est-à-dire celle après retraitements et discrétisation.

En résumé, les bases utilisées en entrée des modèles sont réutilisées sans les variables d'intérêt et avec le genre en variable cible. Les métriques utilisées sont :

1. $$\text{Précision} = \frac{VP + FP}{N}.$$
2. $$\text{F1-score} = \frac{VP}{VP + \frac{1}{2}(FP + FN)}.$$
3. Area Under Curve (AUC) est l'aire sous la courbe Receiver Operating Characteristic (ROC).

Le tableau 4.14 donne les métriques de précision du meilleur de ces modèles.

Modèles	Précision	F1-score	AUC
Logistique	72,56%	68,80%	71,24%
RF	77,61%	70,93%	76,30%
GB	74,04%	70,42%	73,11%

TABLE 4.14 – Performances des modèles de reconstruction du genre après de légères optimisations.

Le niveau de performance obtenu sans effort de modélisation est significatif. Ainsi, comme induit par les résultats de la mesure du biais, les variables explicatives à disposition permettent de reconstruire la variable genre. Par conséquent, bien que cette variable soit omise ou retraitée, sans prise en compte de l'équité, le biais a les moyens de rester présent à travers les autres variables.

En conclusion, un niveau de biais non négligeable est présent dans les données historiques et quelle que soit l'approche de modélisation classique, ce biais est à minima conservé. Ce phénomène est dû aux interdépendances entre la variable sensible et les autres variables explicatives. La figure 4.15 résume l'arbitrage performance-équité entre les différentes approches de modélisation.

Maintenant que la présence du biais a été étudiée, il est temps de passer à la mitigation de ce biais. Dans la section qui suit, cette mitigation est implémentée tout en tentant de maintenir un niveau de performance acceptable. Le modèle construit sans le genre est utilisé comme référence.

4.3 Mitigation du biais

Les modèles ont été construits, les performances et les biais mesurés, il faut maintenant tenter de mitiger le biais tout en préservant le niveau de performances. Dans cette section, diverses méthodes de mitigation sont proposées. La figure 4.16 illustre l'intégration de la mitigation dans le processus de tarification.

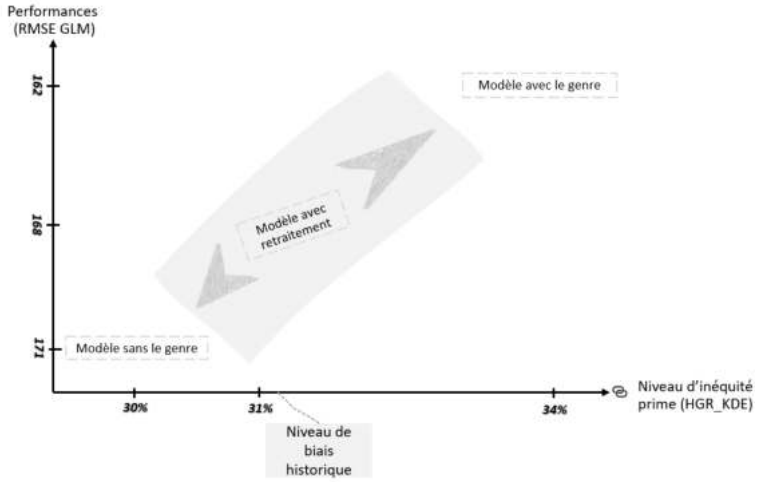


FIGURE 4.15 – Arbitrage entre performance et équité.

FIGURE 4.16 – La place de la mitigation dans le processus de tarification.

4.3.1 Mitigation par le retraitement des données ante modélisation

L'approche de mitigation par retraitement permet de conserver le reste de la chaîne de modélisation tout en permettant le traitement du biais. Les approches de retraitement implémentées sont :

1. Suppression totale ;
2. Suppression de dépendance linéaire ;
3. Adaptation de la méthode fair-SMOTE.

Suppression totale. Cette approche consiste à supprimer des modèles de prédiction toutes les variables ayant une dépendance avec S supérieure à un seuil fixé α. Pour mesurer ces dépendances, le coefficient HGR KDE est utilisé, l'idée étant que les modèles auront plus de difficultés à reconstruire le genre si les variables qui lui sont liées

sont absentes du modèle. La perte d'informations que peut induire cette approche est toutefois un inconvénient non négligeable. Avant l'application de cette méthode, les inter-dépendances sont étudiées. La figure 4.17 donne le niveau de dépendance entre S et les autres variables.

FIGURE 4.17 – Interdépendance entre le genre et les autres variables explicatives.

En observant ces niveaux de dépendance, α peut prendre des valeurs comprises dans l'intervalle $]9\%, 40\%]$. En itérant sur ces différentes valeurs de seuil, des scénarios peuvent être construits, un scénario représentant les variables à exclure de la modélisation. En plus de construire automatiquement les scénarios à l'aide des seuils, ils sont construits sur la base de l'appréciation des liens entre les variables. Les métriques de comparaison des modèles sont la RMSE pour la performance et le HGR KDE pour l'équité. Parmi tous les scénarios testés, les plus intéressants sont les scénarios non-dominés.
Un scénario sera dit non-dominé s'il n'existe aucun autre scénario qui ait à la fois une meilleure performance et moins de biais que lui. Ainsi, les scénarios dominés n'ont pas d'intérêts car quel que soit l'arbitrage que décide de faire l'assureur, il existe un scénario ayant de meilleures métriques. A contrario, un modèle non-dominé peut être le meilleur pour un certain niveau d'équité et ne pas l'être pour un autre.
Les scénarios retenus sont les suivants :

- Scénario 1 : type_boite ;
- Scénario 2 : type_boite et conducteur_2 ;
- Scénario 3 : type_boite, conducteur_2 et energie ;
- Scénario 4 : type_boite, conducteur_2, energie et poids_kw ;
- Scénario 5 : type_boite, conducteur_2, energie et prix_cata ;
- Scénario 6 : type_boite, conducteur_2, energie et zonier ;
- Scénario 7 : type_boite, conducteur_2, energie et anc_cp ;
- Scénario 8 : type_boite, conducteur_2, energie, prix_cata et poids_kw ;
- Scénario 9 : type_boite, conducteur_2, energie, prix_cata et zonier ;
- Scénario 10 : type_boite, conducteur_2, energie, poids_kw et zonier ;
- Scénario 11 : type_boite, conducteur_2, energie, prix_cata, poids_kw et zonier.

La variable genre est retirée de tous les modèles. La figure 4.18 permet de visualiser le niveau d'équité et de performances de chaque scénario. Les scénarios non-dominés sont

en rouge. L'abréviation *tce* représente type_boite, conducteur_2 et energie. Ainsi, le scénario 6 apparaît sur le graphique avec l'abréviation *tce+zonier*.

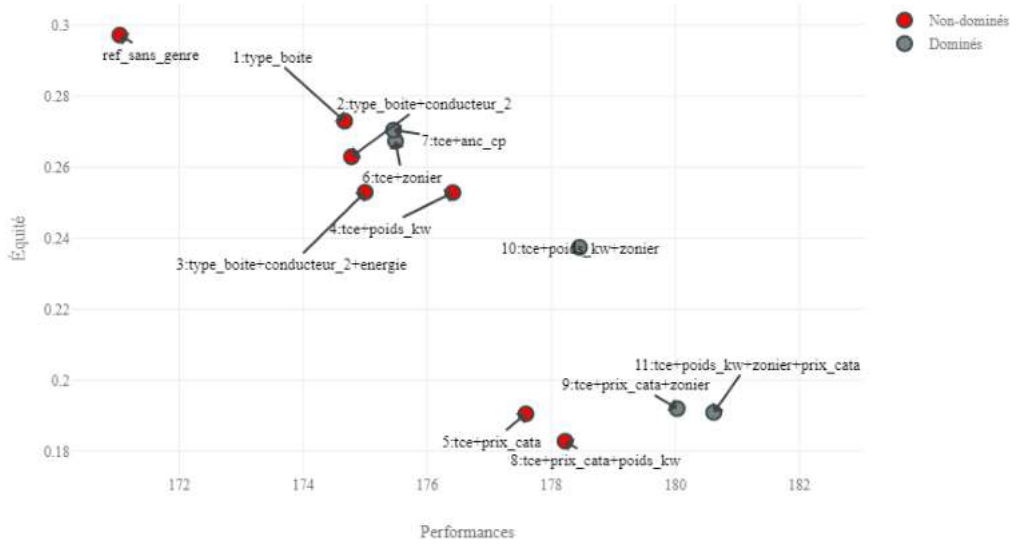
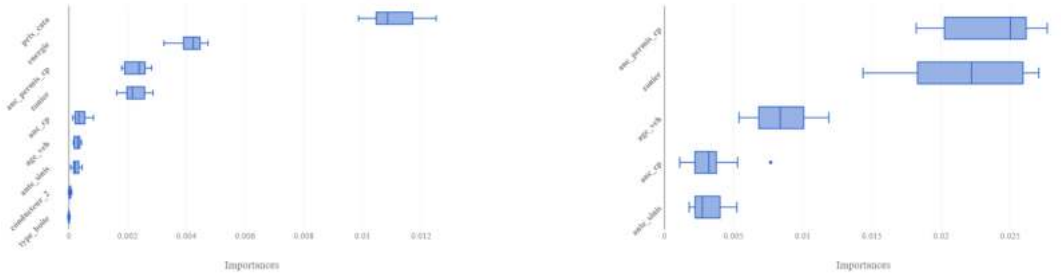


FIGURE 4.18 – Équité des modèles en fonction de leurs performances.

L'observation du scénario 10 permet d'illustrer le principe de scénario dominé. En effet, il a une RMSE de 178,4 et un HGR KDE de 23%. Tandis que le scénario 5 a une RMSE de 177,6 et un HGR KDE de 19%. Le scénario 10 est donc dominé par le scénario 5 car quelle que soit la métrique choisie, le scénario 5 est meilleur.

Parmi les scénarios non-dominés un choix doit être effectué en déterminant l'arbitrage souhaité par les décideurs. Par exemple, le mot d'ordre peut être d'obtenir le modèle le plus équitable possible avec une perte maximale de performance de 5%. Les scénarios 8, 9, 10 et 11 bien que parmi les plus équitables ne seraient plus qualifiables.
En analysant le graphique, le scénario 5 : tce+prix_cata se présente comme un arbitrage intéressant, le biais est quasiment divisé par 2 pour une perte de performance qui semble acceptable. En effet, par rapport à la référence, une baisse du biais de 35% est observée en contrepartie d'une perte de performance de 4%.
Le modèle du scénario 5 est inspecté un peu plus en détails. La figure 4.19 montre l'importance des variables obtenues avec la méthode PFI, ce graphique est comparé à celui qui avait été obtenu avec l'ensemble des variables.
Comme précédemment, pour des raisons de mise à l'échelle, les importances des variables klm_souscrit et poids_kw ont été retirées du graphique PFI GLM. Elles se classent respectivement en première et seconde position en terme d'importance PFI. La hiérarchie des importances est quasiment restée la même avant et après la suppression, mise à part



(a) PFI modèle GLM avant suppression

(b) PFI modèle GLM après suppression

FIGURE 4.19 – Importance avant et après suppression.

age_veh qui est maintenant plus importante que anc_cp.
Il semble que le modèle ait appris à se servir un peu plus des autres variables dans le but de pallier l'absence de la variable prix_cata. Certaines importances sont par exemple deux fois plus grande qu'avant la suppression. La perte de performance montre tout de même que cette variable ne peut pas être optimalement remplacée. De plus, ce nouveau modèle conduit à un équilibre tarifaire très légèrement dégradé par rapport à l'équilibre de référence. Le S/P obtenu est de 99,4% contre le S/P de référence de 99,6%.
Il faut se rappeler qu'une fois qu'un scénario de modification du modèle initial est étudié, une évaluation et une validation du modèle sont de nouveau indispensables.

Ainsi, suivant la qualité des variables explicatives et la perte de performances pouvant être consentie, cette première approche peut conduire à des résultats satisfaisants. Elle est simple à appliquer et permet d'illustrer le fait que la prise en compte des questions d'équité commence par la manière dont les données sont traitées et sélectionnées.

Suppression de dépendance linéaire. L'objectif de cette méthode est de supprimer les relations entre la variable sensible et les variables explicatives tout en conservant assez d'informations pour maintenir les performances des modèles. Pour cela, pour chaque variable explicative X_1, un modèle de régression linéaire est construit pour expliquer X_1 par S centré. La variable corrigée est une combinaison linéaire des anciennes valeurs de X_1 et des résidus de la régression. Le niveau de suppression est contrôlé par l'hyperparamètre α donnant le niveau de régularisation à effectuer sur les données. Celui-ci varie entre 0 et 1 : 0 pour retenir les données initiales et 1 pour retenir uniquement la base de données retraitée. Entre ces deux extrêmes les nouvelles variables explicatives sont une moyenne pondérée des valeurs retraitées et des valeurs initiales.

Dans le cas où X_1 est qualitative, cette approche perd en cohérence : la régression linéaire ne peut plus être utilisée comme modèle. Des modèles de classification ne peuvent non plus être utilisés car la variable sensible est binaire. Le modèle construit ne pourra donc que fournir deux prédictions.
La solution utilisée est de conserver les variables qualitatives et de ne transformer que les

variables quantitatives. Les variables transformées sont : prix_cata, poids_kw, age_veh, anc_cp, anc_permis_cp et klm_souscrit.
La figure 4.20 donne, pour chaque valeur de α, la performance et le biais mesurés dans les modèles sous-jacents construits. Pour générer les valeurs de α, l'intervalle $[0, 1]$ est découpé en 100 parties équidistantes, 100 valeurs de α sont donc obtenues. Sur ce graphique, les modèles non-dominés sont mis en avant en rouge. Le premier constat est qu'aucune

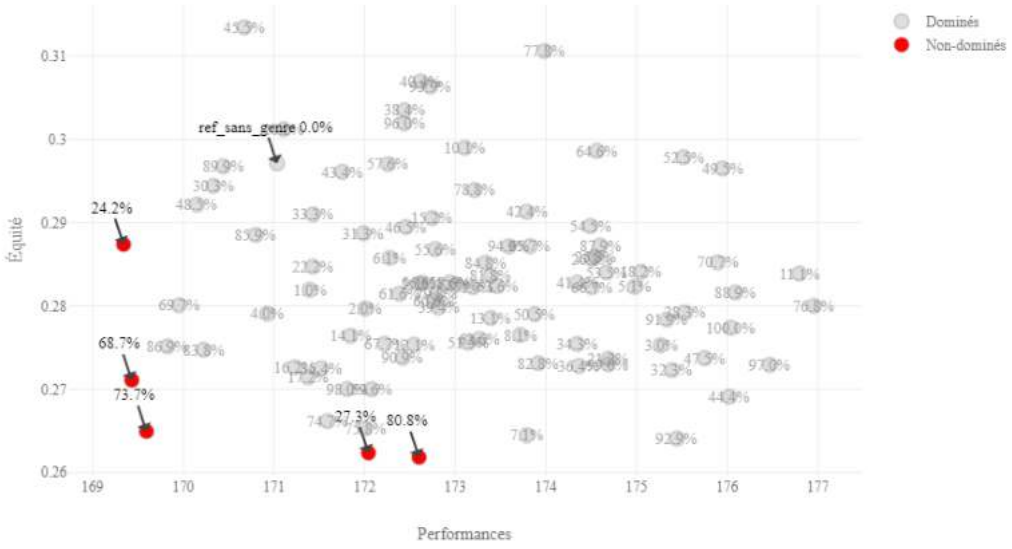


FIGURE 4.20 – Équité des modèles en fonction de leur performance.

tendance ne se dessine entre la dépendance et l'hyperparamètre α. La tendance attendue serait de voir la dépendance et les performances être des fonctions décroissantes de α mais ce n'est pas le cas.
Le second constat est qu'avec cette méthode, le modèle de référence est dominé par certains modèles. En effet, le modèle avec $\alpha = 73{,}7\%$ contient moins de biais et est légèrement plus performant. Le gain de performance est de l'ordre de 0,5% ce qui est négligeable, mais le gain en équité est de 11%. La capacité de suppression de cette méthode dépend de la faculté du modèle linéaire à détecter des relations dans le but que les résidus ne soient plus biaisés. Néanmoins, l'étude des corrélations linéaires montre que la plus forte dépendance entre S et les autres variables explicatives est de 8% ce qui limite fortement la capacité de suppression dans le cas étudié ici.

Cette méthode semble certes intéressante, mais en se replaçant dans le cadre spécifique de la tarification, elle est accompagnée de plusieurs limites. D'abord en tarification, en entrée des modèles, les variables discrètes ou discrétisées sont préférées. Idéalement, il faudrait donc discrétiser les variables continues après suppression. Les valeurs étant des résidus, elles peuvent ne plus être interprétées clairement. De plus, au moment de la

distribution, il faudrait pouvoir transformer les attributs du client pour prédire sa prime. Au vu de ces différents éléments, bien qu'un modèle dominant le modèle de référence ait été obtenu, il est difficile à retenir du fait des différentes limites associées à cette approche. Cette méthode aurait plus de réussite dans le cas d'une variable sensible continue.

Adaptation de la méthode fair-SMOTE. L'approche fair-SMOTE se distingue des deux précédentes dans la mesure où elle ne modifie pas les données existantes mais rajoute des individus artificiels dans la base de données. L'objectif de cette méthode est de s'assurer que les genres soient représentés de la même manière quel que soit le niveau de primes. Ainsi, cette approche peut permettre de pallier les problèmes liés à la sous-représentation de certaines classes. De plus, elle n'est appliquée que sur la base d'apprentissage conservant ainsi intacte la base de test.

Dans le but d'appliquer cette méthode à la variable cible continue, celle-ci est discrétisée permettant ainsi la définition des zones de rééchantillonnage. Il est clair que plus le nombre de zones est grand, plus le découpage permettra d'approcher la distribution continue et donc de réduire le biais statistique provenant de la discrétisation. Toutefois, en choisissant un trop grand nombre de zones, les sous-populations risqueraient de devenir trop faibles pour permettre des simulations cohérentes. Pour les individus générés, la prime continue sera ensuite calculée comme une moyenne pondérée des primes des voisins les plus proches.

Une étude de la distribution des primes prenant en compte ces deux contraintes a conduit à une segmentation en 7 zones. Les zones sont les suivantes :

— zone 1 : $prime = 0$;
— zone 2 : $prime \in]0, 250]$;
— zone 3 : $prime \in]250, 500]$;
— zone 4 : $prime \in]500, 750]$;
— zone 5 : $prime \in]750, 1000]$;
— zone 6 : $prime \in]1000, 1500]$;
— zone 7 : $prime \in]1500, +\infty]$.

Ce nombre de zones peut être considéré comme un hyperparamètre et conduire à la construction de scénarios de segmentation. Ces zones ont été choisies car elles collent plus ou moins bien avec les tendances visibles sur la distribution des primes [1]. Après le choix des zones, le périmètre de rééchantillonnage doit être choisi. En effet, dans la littérature, en plus de rééquilibrer les classes de S, la distribution de Y est aussi rééquilibrée. Cela veux dire que quel que soit Y, il y a autant d'hommes que de femmes et que le nombre total d'individus est aussi le même. L'intérêt de cette démarche est de fournir des données en quantités suffisantes sur toute l'étendue de la distribution de Y.
Cela peut toutefois poser problème en tarification étant donné que la forme de la distribution de Y serait fortement altérée dans la base d'apprentissage. Il y aurait autant d'individus avec des primes de moins de 50€ que d'individus avec des primes de plus de

1. voir figure 4.12b

200€ par exemple. Les lois utilisées pour les GLM ne seraient plus adaptées. Le choix est fait d'équilibrer uniquement les distributions pour le genre, $S|Y$. Pour les algorithmes RF et GB qui ne font aucune hypothèse de loi, le rééchantillonnage sur Y a été testé. Il conduit à un S/P trop faible du fait des niveaux élevés de primes prédits par les modèles. Les auteurs de la méthode fair-SMOTE classique préconise de fixer st et ft à 0,8 chacun. Cette configuration est maintenue. La distribution après rééchantillonnage est donnée par la figure 4.21.

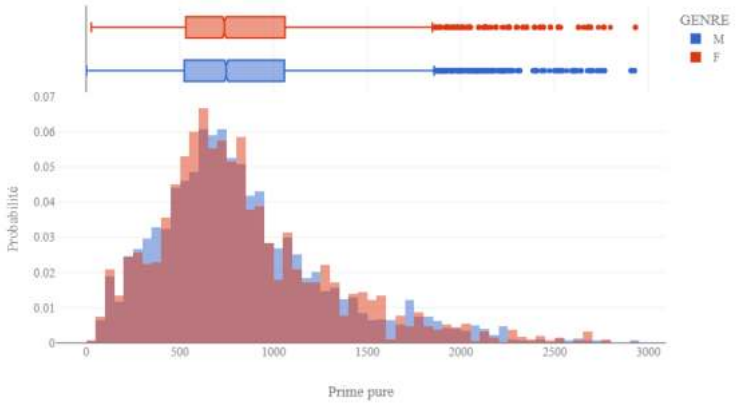


FIGURE 4.21 – Distribution de Y après rééchantillonnage.

Au total, 15488 individus ont été générés, soit une augmentation de la base d'apprentissage de 14,4%. La distribution reste cohérente avec l'historique et les différences entre les genres sont moins visibles. Les boîtes à moustache semblent superposées, le décalage visible dans l'historique est significativement réduit. Ce rééchantillonnage semble donc être concluant. Une fois les données d'apprentissage préparées, les modèles sont de nouveaux construits. Les résultats obtenus sont affichés dans le tableau 4.15.

Modèles	RMSE	HGR_KDE	S/P
Modèle de référence	171,04	29,71%	99,66%
Modèle après fair-SMOTE	171,61	28,83%	99,65%

TABLE 4.15 – Tableau comparatif des résultats après fair-SMOTE.

Bien que la distribution initiale soit moins biaisée, les résultats obtenus sur la base de test sont toujours aussi biaisés. Le gain en équité est inférieur à 0,5%. La performance et le S/P sont aussi restés proches du niveau de référence. Il semblerait donc que rééquilibrer les distributions sur $S|Y$ ne soit pas suffisant pour imposer l'équité. Cela peut s'expliquer par le fait qu'en tarification automobile, le biais à traiter ne provienne pas d'un problème de représentation sur Y mais plutôt d'un écart historique entre les primes des hommes et celles des femmes. Ainsi, puisqu'un rééchantillonnage permet d'assurer que le modèle ait assez de données sur chaque genre pour garantir un traitement égalitaire

dans l'apprentissage, mais ne tient pas compte des interdépendances entre les variables dans ce processus, il ne suffit pas pour garantir l'équité.

Des sensibilités sur les différents choix de modélisation sont effectuées pour vérifier s'ils ne sont la cause de l'échec de la méthode. La modification des hyperparamètres ft et st ne permet pas d'améliorer les résultats. Les zones de rééchantillonnage sont aussi modifiées sans grand succès. L'augmentation du nombre de zones rend la distribution de Y plus irrégulière et réduit la performance.
Une des décisions fortes était de dévier de la littérature et de ne rééchantillonner que sur $S|Y$. Le rééchantillonnage complet sur Y change la distribution de Y, conduit à un trop grand nombre d'individus artificiels et ne produit pas de bons résultats. Pour tester les bienfaits du rééchantillonnage sur Y sans totalement changer la forme de la distribution, les valeurs de Y sont rééquilibrées partiellement. Par exemple, pour un Y qui ne prend que deux valeurs $\{25, 35\}$, les tableaux 4.16 à 4.19 illustre les différentes formes de ré-échantillonnage.

Y/S	F	H	Total
25 €	250	150	400
35 €	120	170	290
Total	370	320	690

TABLE 4.16 – Répartition initiale.

Y/S	F	H	Total
25 €	250	250	500
35 €	170	170	340
Total	420	420	840

TABLE 4.17 – Équilibre sur $S|Y$.

Y/S	F	H	Total
25 €	250	250	500
35 €	250	250	500
Total	500	500	1000

TABLE 4.18 – Équilibre sur $S|Y$ et Y.

Y/S	F	H	Total
25 €	250	250	500
35 €	200	200	400
Total	450	450	900

TABLE 4.19 – Équilibre sur $S|Y$ et partiel sur Y.

Ainsi, plusieurs scénarios de rééchantillonnage partiel de Y sont testés. Le constat est que plus le nombre d'individus simulés s'accroît plus la performance du modèle est dégradée et cela sans que l'équité ne soit meilleure. Par exemple, en accroissant la population de 20% par rééchantillonnage, la RMSE passe à 185 alors que le HGR reste à 28,83%. Au final, aucun de ces différents éléments ne permet d'obtenir de meilleurs modèles.
Une des dernières pistes d'amélioration est de tenir compte des variables explicatives pour le rééchantillonnage. En d'autres termes, au lieu de faire un rééchantillonnage pour avoir l'égalité sur $S|Y$, un rééchantillonnage sur $S|X_1, \dots X_M, Y$ peut être effectué. Dans cette approche, le choix des variables pour le conditionnement est crucial, ces variables pouvant conduire à une amplification des biais déjà existants. De plus, son implémentation est délicate car il faut contrôler la qualité du rééchantillonnage sur les variables prises individuellement mais aussi sur les croisements, le tout en conservant la cohérence globale des primes. Cette piste n'a pas été explorée plus en détails, mais des méthodes plus efficientes dans la prise en compte des interdépendances peuvent être intégrées à

l'avenir.

Ces différentes approches de mitigation ante modélisation ont permis de découvrir plus en détails la structure des données et les éléments ayant une influence significative sur les niveaux de biais obtenus. Elles fournissent des outils permettant la prise en compte des enjeux éthiques dans la phase de traitement des données.

4.3.2 Mitigation pendant la modélisation

Les méthodes de mitigation pendant l'étape de calibrage des modèles permettent d'introduire directement les contraintes d'équité dans la résolution du problème d'optimisation. Deux méthodes de cette famille sont implémentées et testées. Ce sont l'exponentiated gradient et sa version recherche par grille. Pour appliquer ces méthodes à une définition quelconque de l'équité, il est nécessaire de réécrire les contraintes associées à cette définition sous forme de systèmes d'inéquations [2]. Ces méthodes étant initialement appliquées à la mitigation du biais dans le cadre de la classification, les contraintes de définition de parité statistique et de chances égalisées sont disponibles dans la littérature. En théorie, les mêmes travaux peuvent être menés dans le cas continue. Toutefois, aucune réécriture des définitions classiques n'est disponible. Les seuls travaux accessibles utilisent, comme définition de l'équité, l'égalité des performances des modèles sur chaque classe de la variable sensible. La contrainte prend la forme suivante :

$$\mathbb{E}(L(Y, \widehat{Y})|S = s) < \zeta, \forall s.$$

Cette contrainte permet de s'assurer que le modèle commet en moyenne les mêmes erreurs quel que soit le genre. En d'autres termes, les prédictions sont aussi bonnes pour les hommes que pour les femmes. Bien que cette contrainte soit intéressante, elle est plus égalitaire qu'équitable. Comme pour le fair-SMOTE, elle ne permet pas de rééquilibrer les niveaux de primes entre les genres.
Dans le but de pouvoir appliquer les définitions principales de l'équité au cas continu des primes, des tentatives de réécriture des contraintes ont été menées sans succès ; ni mathématiquement, ni algorithmiquement ces contraintes n'ont pu être implémentées pour l'exponentiated gradient. Il se pourrait que ces contraintes ne puissent pas être appliquées telles quelles avec cette méthode. Faute de temps, cette piste n'a pas pu être explorée plus longtemps pour soit trouver une solution soit prouver l'existence ou non d'une solution.
Les résultats présentés sont donc ceux obtenus avec la contrainte de majoration des erreurs de prédictions. Le paramétrage des méthodes réside donc dans celui de la fonction de perte. Deux fonctions de pertes sont testées, la MAE et la RMSE telles que définies pour mesurer les performances des modèles. Le paramètre ζ permet de préciser le niveau jusqu'auquel les contraintes peuvent être enfreintes. En plus de ce paramètre, il est possible de borner l'erreur provenant d'une prédiction. Pour cela la relation se réécrit :

$$\mathbb{E}[\max(L(Y, \widehat{Y})|S = s, M)] < \zeta, \forall s.$$

2. Voir la section 3.2.1

Ainsi, lorsque l'algorithme ne parvient pas à converger, le seuil peut être rajouté pour limiter les écarts pris en compte.
L'approche adoptée est de relancer l'algorithme avec un ζ le plus faible possible qui est ensuite augmenté au fur à mesure jusqu'à ce que la méthode converge. Donc en partant de $\zeta = 10^{-5}$ et en augmentant par multiple de 10, la première convergence a eu lieu pour $\zeta = 10^{-1}$. Une fois que la méthode converge vers la solution optimale, la recherche par grille et le paramètre M ne sont plus utiles.
Par contre, pour les $\zeta < 10^{-1}$, ils peuvent permettre de trouver des solutions suboptimales qui pourraient être meilleures que la solution trouvée avec des tolérances plus grandes. Pour les tests, des valeurs de M de 0.5 à 500 sont utilisées avec à chaque fois des grilles de taille 3000. Les résultats obtenus se sont montrés moins bons que ceux obtenus avec la convergence. Il est toutefois important de préciser que les grilles explorées étaient de petites tailles par rapport à la dimensionnalité du jeu de données. Avec plus de puissance et de temps de calcul, de meilleurs résultats auraient potentiellement pu être obtenus.
Les résultats obtenus pour le meilleur modèle sont présentés dans le tableau 4.20.

Modèles	RMSE	HGR_KDE	S/P
Modèle de référence	171,04	29,71%	99,66%
Modèle après mitigation	171,25	31,22%	99,65%

TABLE 4.20 – Évaluation des résultats obtenus après application de l'exponentiated gradient.

Le modèle obtenu après mitigation est moins performant et moins équitable que le modèle de référence. La perte de performance est négligeable mais le modèle a un HGR plus grand de 7% après l'application d'une méthode de mitigation. Ces résultats s'expliquent par le fait que la contrainte d'erreur implémentée n'impose aucune forme d'indépendance entre $\widehat{Y}$ et S. En effet, le taux d'erreur par genre peut être le même sans que les genres ne soient traités de manière équitable.

En dehors du fait que les autres définitions de l'équité n'ont pas pu être implémentées, cette méthode a des temps de calculs exponentiels. La méthode de recherche par grille et le paramètre M fournissent toutefois une alternative, mais une alternative qui conduit à des résultats suboptimaux. Dans ce domaine en pleine croissance qu'est la mitigation des biais, la mitigation pendant la modélisation est l'aspect le plus délicat et donc le moins développé. Les quelques méthodes accessibles sont axées sur la classification binaire et la grande personnalisation des méthodes rend leur adaptation difficile. Toutefois, la méthode étudiée ici, implémentée sous une contrainte relativement simple d'équité permet d'illustrer les limites de la mise en place de l'équité comme contrainte d'optimisation. Comme abordé dans la section "le coût de l'équité", l'implémentation de ces contraintes conduit à des temps de calculs exponentiels même dans les cas les plus simples. De plus, l'obtention de méthodes accessibles, généralisables et stables est un réel défi qui doit être abordé du point de vue mathématique en prenant en compte les contraintes qu'imposent la tarification.

4.3.3 Mitigation post modélisation

La construction de frontières de décisions dans la classification binaire est un aspect exploité dans la littérature pour construire les méthodes de mitigation post modélisation. L'exploitation des probabilités, du seuil de décision et des métriques telles que le rappel et la précision, dans le cas de la régression logistique, donne une illustration concrète de l'implémentation d'une mitigation post modélisation. Le seuil est déplacé tout en observant les effets sur les différents genres et une fois l'équilibre souhaité entre équité et performance trouvé, les prédictions peuvent être réévaluées en utilisant les probabilités sans devoir reconstruire le modèle.
Dans le cas de la régression, aucune publication appliquant et testant la mitigation post modélisation n'a pu être trouvée. Aussi, les principes utilisés dans le cas de la classification ne sont pas applicables. L'approche "Redistribution équitable" est donc proposée dans le but d'imposer une équité individuelle post modélisation.
Pour pouvoir faire la proposition d'une méthode imposant l'équité de groupe, il faudrait pouvoir définir ce qu'est une prime avantageuse et une prime désavantageuse. Pour l'assuré, la prime avantageuse est certainement la prime la plus faible ; qui ne voudrait pas payer moins ! Plus sérieusement, il n'est pas possible de définir dans l'absolu la notion de prime avantageuse car celle-ci dépend des caractéristiques de l'assuré et du risque porté. C'est à la suite de ce constat que l'approche individuelle a été proposée.

Redistribution équitable. L'adaptation du flip-test permet de mesurer le biais individuel, c'est-à-dire le biais subit par chaque individu de la base de données. Pour éviter d'entrer dans les statistiques causales, des algorithmes de k plus proches voisins sont utilisés pour détecter les individus semblables mais de genre différents. Cette approche permet d'estimer pour une femme donnée, l'écart moyen de primes observé avec les hommes qui lui sont le plus proche. Le même raisonnement peut être reconduit en remplaçant femme par homme. Pour les prédictions $\widehat{Y}$ obtenues avec le modèle GLM retenu pour la prime, la distribution des écarts pour les femmes est visible sur la figure 4.22. Elle a la forme d'une gaussienne et est de moyenne $-0{,}9$€ car il y a un peu plus d'observations dans la partie négative.
Le modèle de k plus proches voisins a fait l'objet d'un hyperparamétrage et d'une sélection de variables optimales avec comme objectif d'obtenir les plus petits écarts moyens possible sur une base de test. En d'autres termes, en plus de contrôler la qualité du modèle avec la notion de distance présente dans le k plus proches voisins, les écarts de primes entre les femmes et leur voisinage sont minimisés.
Les variables retenues sont : anc_cp, klm_souscrit, conducteur_2, age_veh, prix_cata, type_boite et zonier pour un nombre de voisins de 6 et une distance de Manhattan. Sur le langage de programmation Python, un attribut de la méthode de k plus proches voisins permet de laisser l'algorithme sélectionner la méthode optimale de recherche de voisins en se basant sur la dimensionnalité, le nombre d'individus et la structure des données (par exemple si la matrice est creuse ou non). Cette méthode automatique a permis de trouver les meilleurs résultats.

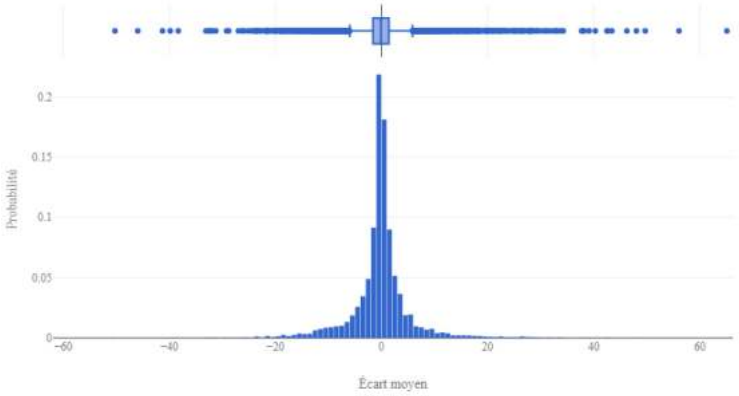


FIGURE 4.22 – Histogramme écarts moyens entre les femmes et leur voisinage masculin.

Au final, pour chaque individu de la base de données, l'écart moyen avec ses voisins les plus proches de genre opposé est calculé. Ces écarts moyens sont joints à la base de modélisation initiale. L'expérience suivante est menée sur une base de test pour des totaux de charges de 411.296€ et de primes prédites de 413.202€ sur 16000 observations. Le S/P de 99,5% est cohérent avec le S/P global de 99,6%. En moyenne, les primes des femmes sont plus faibles de 0,9€ par rapport à celles des hommes les plus proches. Une étude détaillée des distributions permet néanmoins de constater que 30% des femmes paient une prime 15% moins chère que celle des hommes voisins. Le tableau 4.21 résume les agrégats sur le périmètre de test.

Agrégats	Femme	Homme	Somme
Charges totales	153238€	258057€	411296€
Primes prédites	154953€	258248€	413202€
Exposition	5459	8131	13590
Nombre d'individus	6937	9063	16000
Écart moyen	-0,9€	1,2€	0,3€
Somme des écarts	-25535€	32019€	6484€

TABLE 4.21 – Quelques agrégats pertinents sur le périmètre de test.

La somme des écarts pour les femmes est de -25535€ contre $+32019$€ pour les hommes. Ces écarts s'interprètent de la façon suivante : au global, les femmes paient 25535€ de moins que les hommes qui leur sont semblables. Les signes sont cohérents car les femmes portent moins de risques que les hommes. Toutefois, les écarts ne se compensent pas parfaitement : $-25535+32019=6484$€. Cet écart résiduel provient en partie de la présence de plus d'hommes dans la base mais aussi du fait que les modèles utilisés ne soient pas parfaits.

L'idée ici est qu'il faut corriger les primes dans le but de réduire les écarts entre hommes et femmes. Les écarts seraient alors recalculés sur les primes corrigées pour vérifier l'équité du nouvel état obtenu. La qualité des nouvelles primes doit aussi être étudiée.
Intuitivement, la correction des primes peut se faire de la manière suivante : $prime\ i = prime\ i - ecart\ i$. Toutefois, cela reviendrai à déplacer le problème car les femmes auraient maintenant des primes plus élevées que les hommes. En effet, les primes hommes sont baissées et les primes femmes sont augmentées, il y a donc double correction. Dès lors, faut-t-il se retrouver à mi-chemin en divisant l'écart par 2 ?

$$prime\ i = prime\ i - \frac{ecart\ i}{2}.$$

Cela aurait pu fonctionner dans de petites dimensionnalités. Le problème avec cette approche est qu'en rapprochant la prime d'un individu de la prime moyenne de ses voisins de genre opposé, elle s'éloigne potentiellement des primes d'autres individus dont il composait le voisinage. Dans le cas d'usage étudié, la correction de l'intégralité des écarts a conduit à une augmentation des disparités au global. Pour réussir à rapprocher la prime d'un individu des primes de son voisinage sans l'éloigner des primes des individus dont il constitue le voisinage, une approche récursive est proposée. Elle se détaille de la manière suivante :
Commencer par mesurer l'écart sur un premier genre $s = 0$ puis, tant que la somme des écarts entre les genres est supérieur à un seuil ϵ fixé, répéter les quatre étapes suivantes :

1. Ajuster les primes pour le genre s :

$$prime\ i = prime\ i - \frac{ecart\ i}{\eta}, \quad \forall i \text{ tel que } S_i = s.$$

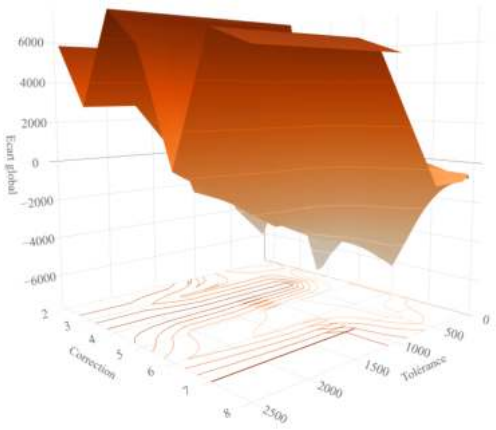
2. Prendre le genre opposé, c'est-à-dire si $s = 0$, prendre $s = 1$ et vice versa.
3. Mesurer l'écart entre les individus de genre s et leurs voisinages de genre opposé :

$$ecart\ i = prime\ i - prime\_moyenne_{V(\bar{s})}\ i, \quad \forall i \text{ tel que } S_i = s.$$

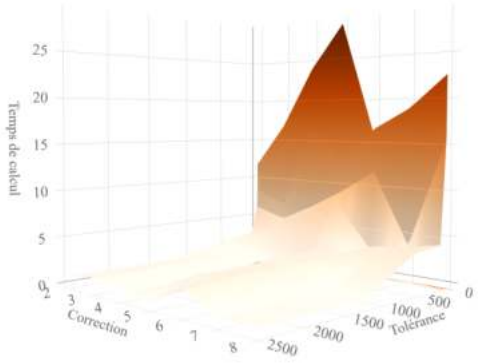
4. Calculer la somme des écarts : $\sum_i ecart\ i$.

Dans le but de mesurer l'effet du choix de ϵ et η et de trouver la combinaison optimale, la redistribution est réitérée sur une grille de valeur. Cette grille est constituée de toutes les combinaisons (η, ϵ) telles que $\epsilon \in \{2000, 1000, 500, 100, 10, 1, 0.1, 0.01, 0.001\}$ et $\eta \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Trois valeurs sont observées : le temps de calcul, l'écart global et la fidélité de la redistribution. L'écart global est l'écart entre le total de prime avant et après redistribution. La fidélité est représentée par le ratio :

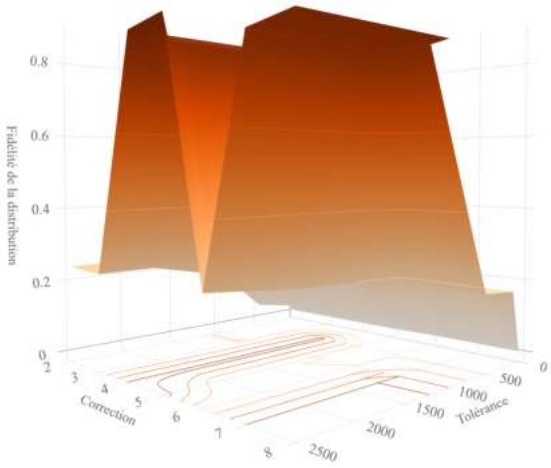
$$\frac{(\max - \min)\text{des primes après redistribution}}{(\max - \min)\text{des primes avant redistribution}}.$$



(a) Écart global en fonction de la tolérance et du facteur de correction.



(b) Temps de calculs en fonction de la tolérance et du facteur de correction.



(c) Fidélité en fonction de la tolérance et du facteur de correction.

FIGURE 4.23 – Arbitrage fidélité, temps de calculs et écart global.

Elle permet de s'assurer que toute l'étendue de la distribution de $\widehat{Y}$ et la diversité des primes prédites sont conservées. La figure 4.23 permet de visualiser les résultats.

Le graphique sur la fidélité permet de constater qu'à partir de $\eta <= 100$, la redistribution conduit à une réduction significative de l'étendue de la distribution de $\widehat{Y}$. En effet, le nombre de corrections à effectuer pour atteindre ces niveaux d'écarts est si grand que les primes se regroupent de plus en plus jusqu'à converger vers la prime moyenne de l'échantillon. Ainsi, en faisant converger le plus possible la méthode de redistribution, un équilibre dans lequel tous les individus ont la même prime est atteint. Cet équilibre est un équilibre équitable trivial, puisqu'aucune distinction ne peut exister entre les genres mais les performances du modèle sont détruites. Il faut donc réussir à faire converger la méthode vers des optimaux où l'écart global est réduit et la distribution $\widehat{Y}$ est conservée.

Sur les 70 combinaisons testées, seulement 9 d'entre elles sont non-dominées en utilisant la fidélité et l'écart global comme critères. Sur ces 9, 5 ont une fidélité de moins de 25%, elles ne sont donc pas acceptables bien qu'elles soient les combinaisons ayant les plus petits écarts. Les deux meilleurs scénarios sont :

1. $\eta = 5$ et $\epsilon = 1000$ pour un écart global de 1297€ et une fidélité de 87% ;
2. $\eta = 8$ et $\epsilon = 2000$ pour un écart global de 1146€ et une fidélité de 83%.

Le scénario avec 87% de fidélité est préféré ; il conduit à une somme des écarts de -850€ pour les femmes et de $+2445$€ pour les hommes. Ainsi, $-850 + 2245 = 1395$€ soit une baisse de 78% par rapport à l'écart initial de 6484€.
Les écarts moyens sont maintenant de 0,075€ pour les femmes et de 0,13€ pour les hommes, tout ceci en conservant une distribution sur $\widehat{Y}$ fidèle à celle avant la redistribution. La figure 4.24 présente la superposition des histogrammes avant et après redistribution.
En ce qui concerne les performances, les primes après redistribution ont une RMSE de 171,66 ce qui est très proche du niveau de référence de 171,04. Avec l'augmentation de la prime de 1297€, le S/P passe de 99,5% à 99,2%. Les primes restent donc cohérentes et bien calibrées tout en réduisant les écarts entre homme et femme.
Pour $\eta <= 100$, les temps de calculs sont pour la plupart inférieurs à cinq minutes, ce critère n'est donc pas pris en compte dans la définition des meilleures redistributions.

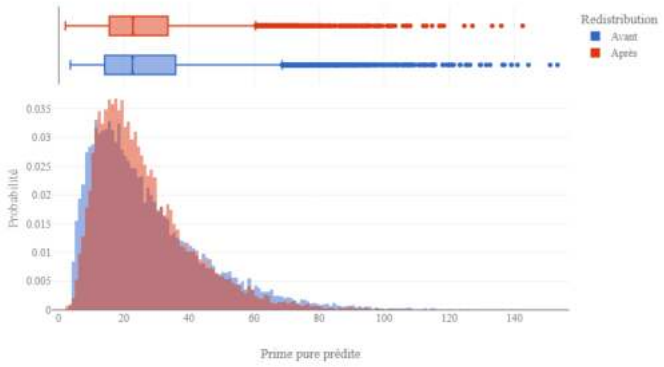


FIGURE 4.24 – Distribution de $\widehat{Y}$ avant et après redistribution.

Pour que cette méthode soit viable, il faut qu'elle puisse être facilement utilisable à grande échelle. Comme discuté dans la présentation formelle de la méthode, il y a deux possibilités : soit construire une grille contenant tous les ajustements possibles soit construire un modèle permettant de prédire les ajustements. La première solution est simple à implémenter mais peut être difficile à distribuer. La seconde quant à elle, accroît le risque de biais de modélisation mais est plus simple à distribuer et peut servir

à corriger les coefficients GLM du modèle de prime. La modélisation est implémentée dans le but de fournir une évaluation rapide de son potentiel.

Modélisation des écarts. Pour la modélisation de ces écarts, les modèles linéaires sont choisis car ils sont simples, facilement interprétables et leurs coefficients peuvent directement servir pour la distribution. Pour replanter le décor, des primes ont été calculées avec un modèle GLM non équitable. Ensuite l'adaptation flip-test a permis de mesurer les biais présents dans ces primes. Ainsi, en utilisant la redistribution équitable, un montant de correction a été calculé pour chaque prime de telle sorte que l'équité individuelle soit la plus forte possible. La prime finale est donc :

$$prime_{finale}\ i = prime_{initiale}\ i - ecart_{final}\ i.$$

L'objectif est de modéliser, à l'aide des variables explicatives à disposition, les écarts pour ne plus avoir besoin de grille ou du modèle de k plus proches voisins. Comme le montre la figure 4.22, la distribution de ces écarts semble proche de la distribution d'une gaussienne. Des modèles linéaires standards sont donc testés. Bien qu'ils conservent l'équilibre tarifaire et ont des performances acceptables, les coefficients obtenus sont grands, de l'ordre de 10^2. La sélection des variables permet de réduire ce problème sans toutefois le résoudre.
Un modèle pénalisé Lasso est donc implémenté. Il permet de faire tendre le plus possible la valeur des coefficients vers 0. Le paramètre de pénalisation est optimisé sans faire de sélection de variables, des valeurs allant de 1 à 0,01 sont testées. Les meilleurs résultats sont obtenus pour un facteur de 0,1. La RMSE obtenue est 9,21 ce qui est non négligeable mais acceptable connaissant l'étendue des écarts. Les variables prix_cata, conducteur_2 et zonier sont les variables ayant les plus grands coefficients non nuls.
Cette méthode peut être utile à condition d'étudier rigoureusement ses effets sur le modèle de prime initiale. L'utilisation de modèle linéaire permet d'assurer le maintien des équilibres moyens et l'obtention de coefficients pouvant être utilisés pour la distribution. Les risques de biais de modélisation sont toutefois non négligeables. Si l'acteur en a les moyens la distribution d'une grille d'ajustement reste l'approche la plus fiable.

Remarques : Cette démarche de modélisation est différente des approches de modélisation de résidus utilisées pour les zoniers par exemple. En effet, les écarts ne représentent pas des erreurs de prédictions mais plutôt la partie biaisée de la prime :

$$prime_{biais}\ i = prime_{equitable}\ i + ecart_{final}\ i.$$

La réutilisation des variables explicatives pour la modélisation d'une partie de la prime est donc cohérente car elles permettent déjà de modéliser toute la prime.

La mise en place de la mitigation a conduit à la prise en compte de l'équité à différentes étapes du processus de tarification. L'intégration de ces mitigations requiert qu'elles soient cohérentes avec les différents besoins et contraintes de tarification. Elles doivent

donc être opérationnellement abordables tout en préservant au maximum la performance des modèles. Parmi les méthodes implémentées, la suppression des variables par sa simplicité et la redistribution par son approche individuelle semblent être les plus concluantes. Les autres méthodes permettent de compléter la compréhension du biais étudié et pourraient conduire à de meilleurs résultats sur d'autres données. Les résultats des différentes mitigations implémentées sont compilés dans le tableau 4.22.

Métriques	Modèle de référence	Suppression totale	Suppression corrélation	Adaptation fair-SMOTE	Exponentiated gradient	Redistribution équitable
HGR KDE	29,71%	19%	26,50%	28,83%	31,22%	30,08%
RMSE	171,04	177,6	169,6	171,61	171,25	171,66
S/P	99,66%	99,30%	99,82%	99,65%	99,65%	99,20%

TABLE 4.22 – Récapitulatif des résultats des mitigations du biais.

Conclusion

Le point de départ de cette étude est la prise en compte des différentes contraintes qui entourent une tarification, celle-ci se devant d'être à la fois compétitive, statistiquement cohérente, transparente et équitable. L'équité bien qu'importante du point de vue social et règlementaire n'est pas encore assez traité en science actuarielle. L'assureur peut, pour s'aligner à la réglementation ou à sa stratégie, devoir proposer des primes plus équitables par rapport à des variables dites sensibles.

C'est dans le but de répondre à ce besoin que les définitions et les mesures de biais éthiques ont été introduites. Pour cela, il a fallu une revue extensive d'une littérature encore en construction dans le but de comprendre, cerner et présenter tous les éléments utiles pour ces travaux actuariels et les suivants. Des propositions et adaptations ont aussi été formulées pour mieux prendre en compte le cas continu. Suivant la problématique et les données à disposition, certaines mesures sont plus cohérentes que d'autres. Six d'entre elles ont été choisies pour le cas pratique traité, celles-ci se complétant et permettant de cerner les différents aspects du biais.
Le coefficient HGR KDE et l'adaptation du flip-test ont joué un rôle prépondérant dans la mesure des biais. Ces mesures ont permis de détecter du biais dans les données historiques mais aussi dans les prédictions des modèles qu'ils contiennent le genre, le suppriment ou le retraitent.
Les biais présents dans les données historiques ont donc été appris et dans certains cas amplifiés par la modélisation. Les interdépendances entre le genre et les autres variables explicatives ont permis de conserver l'effet du genre même quand celui-ci était retiré ou retraité. Les approches classiques montrent ainsi leur limite dans la mise en place de l'équité. D'où la nécessité d'introduire des méthodes permettant une meilleure prise en compte de l'équité tout en conservant la performance et la cohérence des modèles construits.

C'est dans le but de traiter ces biais que les méthodes de mitigation ont été étudiées. La littérature sur ce sujet est, comme dans le cas des mesures, principalement tournée vers la classification. De plus, les méthodes proposées dans la littérature sont souvent trop rattachées au cas pratique traité pour être généralisées. Toutefois, des solutions ont pu être implémentées avec plus ou moins de succès. Des approches simples de mitigation ante modélisation telles que la suppression de corrélations linéaires et de variables liées

au genre ont permis d'obtenir des résultats dont la qualité dépend des contraintes fixées par l'assureur.
L'adaptation du fair-SMOTE a produit une méthode permettant de traiter efficacement les problèmes de représentation dans le cas continu. Elle n'a toutefois pas été efficace sur le problème traité car le biais de représentation était relativement faible.
L'exponentiated gradient et sa méthode de recherche par grille ont connu la même fin. En effet, même s'ils ont permis d'implémenter une mitigation pendant la modélisation, la contrainte d'égalité des erreurs ayant pu être utilisée ne suffisait pas pour imposer une équité assez stricte. Ces méthodes ont montré les difficultés qui entourent la mitigation pendant la phase de modélisation : les méthodes sont spécifiques et les temps de calcul sont longs.
A travers ces méthodes, l'objectif était de mitiger le biais de groupe tel que mesuré par le HGR KDE. La redistribution équitable a ensuite été proposée dans le but d'imposer une équité individuelle post modélisation en s'appuyant sur la définition du biais fournie par l'adaptation du flip-test. Cette méthode a permis de réduire les écarts de primes qui existaient entre les individus et leurs voisinages de genre opposé tout en conservant en grande partie la cohérence des primes.
A chaque étape de mitigation, des hyperparamètres et des scénarios ont été définis dans le but d'optimiser et de rendre personnalisables les méthodes utilisées.

Dans le but de généraliser les résultats des mesures et mitigation, plusieurs jeux de données et de garanties pourront être étudiés, des mailles différentes de tarification pourront aussi être envisagées telles que la maille portefeuille ou groupe de garanties dans le but de venir compléter ces premiers résultats sur l'équité en tarification.
Ces travaux offrent un cadre pour la mesure et la mitigation du biais. L'ensemble des méthodes implémentées couvre un spectre assez large de l'équité ce qui permet de comprendre la forme d'équité à appliquer au cas de tarification traité et de fournir des méthodes adaptables à chaque problématique. Dans la continuité de cette étude, des approches plus efficientes de mitigation pendant la modélisation peuvent être implémentées en modifiant les fonctions objectives de certains algorithmes, des réseaux neuronaux plus flexibles peuvent être utilisés pour mettre en place l'équité à plusieurs niveaux du processus de tarification.
Un autre aspect intéressant de l'équité est celui de la prise en compte de plusieurs variables sensibles à la fois. L'utilisation massive des données personnelles risque d'entraîner à l'avenir de nombreuses évolutions de la règlementation. Il serait intéressant de construire des outils permettant d'implémenter l'équité sur plusieurs variables sensibles et de mesurer l'impact de ces implémentations sur les performances des modèles de tarification.

Annexe A

Interprétabilité des modèles d'apprentissage

Cette partie permettra de présenter l'interprétabilité, d'acquérir les outils nécessaires à l'interprétation des modèles.

A.1 Modèles nativement interprétable

Dans cette section le modèle GLM est présenté comme modèle de référence en assurance et l'arbre de décisions comme le modèle de base de méthodes plus complexe tel que le gradient boosting ou la forêt aléatoire.

A.1.1 Modèle GLM

En reprenant les notations suivantes :
- g la fonction de lien ;
- Y_i la variable cible ;
- X_i la matrice des variables explicatives, où chaque colonne est une variable explicative ;
- β le vecteur des paramètres à estimer.

Une réalisation donnée se réécrit :

$$g(Y_i) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

$$Y_i = g^{-1}(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)$$

Dans les cas usuelles :
$g = Id$,

$$Y_i = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

la prédiction Y se décompose par définition comme une combinaison linéaire des variables X_i.

$g = log$,

$$Y_i = exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)$$

$$Y_i = exp(\beta_0) \times exp(\beta_1 X_1) \times ... \times exp(\beta_n X_n)$$

Alors la prédiction se décompose comme le produit des effets individuels de chaque variable.

A.1.2 Arbre de décision

Les arbres de décisions sont par définition interprétables du fait que le modèle final obtenu sépare les données à chaque nœud en utilisant des règles de décisions sur les variables. La figure A.1 en fourni une illustration simple tirée de travaux d'Olivier Lopez en 2015.

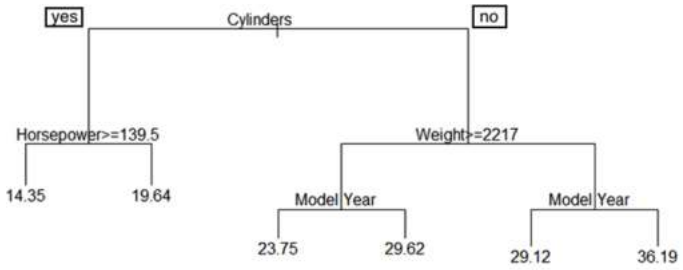


FIGURE A.1 – Illustration arbre de décision.

Ainsi, un fois une prédiction obtenue, il suffit de remonter l'arbre en lisant les règles de décisions pour comprendre comment la prédiction a été obtenue.

En dehors des modèles présentés plus haut, la majorité des modèles de machines learning ne sont pas nativement interprétables. Il n'est pas possible d'avoir accès à des coefficients, une fonction de décision ou des règles de décisions. Dans la plupart des cas, il est difficile de construire des interprétations basées sur le comportement propre aux modèles. C'est pour cette raison que par la suite des méthodes d'interprétabilité agnostiques aux modèles sont présentées. Ces méthodes sont flexibles. En effet, elles permettent d'interpréter un modèle quel qu'il soit d'où le terme "agnostique". Ainsi, il est possible d'utiliser une grande quantité de modèles performants sachant qu'ils pourront être interpréter avec les mêmes outils. Ces méthodes s'opposent aux méthodes d'interprétabilité propres aux modèles, par exemple les coefficients du GLM.

Ces méthodes agnostiques coûtent généralement beaucoup plus chères en temps de calcul. Les interprétations propres aux modèles étant générées au cours du processus de construction du modèle. Tandis que les méthodes agnostiques sont ajustées après implémentation des modèles et requièrent souvent d'effectuer un grand nombre de prédictions. Il existe de nombreuses méthodes d'interprétabilité agnostique.

A.2 Méthodes d'interprétabilité agnostique globale

A.2.1 Permutation Feature Importance (PFI)

La méthode PFI est une méthode qui permet d'avoir une vue d'ensemble sur le rôle joué par les variables explicatives dans le processus de décision. Cette méthode consiste à mesurer l'effet de la permutation des valeurs d'une variable explicative sur le niveau d'erreur du modèle. En notant : $\widehat{f}$ le modèle entraîné, X la matrice contenant en colonnes les p variables explicatives, y le vecteur réponse et $L(.)$ la fonction de perte ; la méthode PFI peut se décrire de la manière suivante :

- Calcul de l'erreur de prédiction du modèle initial, $\epsilon_0 = L(\widehat{f}(X), y)$.
- Pour chaque variable j, $j \in 1, ..., p$ d'abord, générer une nouvelle matrice $X_j^{'}$ en permutant les valeurs de la colonne j. Puis, calculer la nouvelle erreur : $\epsilon_j^{'} = L(\widehat{f}(X_j^{'}), y)$. Enfin déduire, l'importance Imp_j en faisant : $Imp_j = \epsilon_j^{'} - \epsilon_0$ ou $Imp_j = \frac{\epsilon_j^{'}}{\epsilon_0}$.
- Ranger les Imp_j dans l'ordre croissant pour obtenir les variables les plus importantes.

L'idée de cette méthode a été introduite par Breiman[10] (2001) pour mesurer l'importance des variables dans un modèle de forêts aléatoires. En 2018, Fisher, Rudin et Dominici[25] proposent la version agnostique explicitée plus haut. L'idée avancée par cette méthode est assez intuitive. Pour une variable fixée, la permutation de ces valeurs rompt la relation entre la variable en question et y. Ainsi, si le modèle réussit à maintenir son niveau de performance, la variable peut être considérée comme étant sans importance pour la prédiction. Dans le cas où la permutation conduit à de grandes erreurs, la variable sera considérée comme importante.
Cette méthode permet d'avoir une vue globale du comportement du modèle sans avoir à le réentrainer. En outre, en utilisant $Imp_j = \frac{\epsilon_j^{'}}{\epsilon_0}$, PFI permet de produire un indicateur d'importance comparable entre différents modèles.
De plus, la permutation des valeurs permet de prendre directement en compte les interactions entre les différentes variables. Ainsi, l'importance calculée contient l'effet de la variable sur la variable cible mais aussi celui de l'interaction. Cela constitue aussi un inconvénient car dans les modèles avec interactions, l'importance des variables n'est pas égale à la baisse totale de performance. Cela implique que les effets dus aux interactions entre variables sont pris en compte plusieurs fois. De plus, en présence de multicolinéarité, les permutations peuvent donner lieu à des instances de données irréalistes. Le fait de ne pas réentrainer le modèle permet de réduire le temps de calcul de cette méthode. Néanmoins, l'aléa introduit par les permutations peut fortement faire varier les résultats entre deux exécutions. La solution serait de répéter plusieurs fois la procédure énoncée plus haut et de moyenner les importances relatives obtenues. Cette solution augmente hélas significativement le temps de calcul.

A.2.2 ALE

La seconde méthode présentée est la méthode ALE. Elle permet aussi de mesurer l'effet d'une variable sur le modèle. La méthode présentée précédemment a fait ressortir deux principaux problèmes : la présence d'instances irréalistes et la prise en compte des interactions dans la mesure de l'importance. La méthode ALE permet de palier à ces inconvénients. Elle est une version améliorée des méthodes PDP (Partial Dependence Plot) et M-plot (conditional plot). L'idée originelle qui est celle du PDP est que pour une variable fixée X_j, la démarche suivante est adoptée :

- Définition d'une grille de valeurs ($\mathbb{G}$) à tester (des variations des valeurs de X_j observées dans le jeu de données).
- Pour chaque $v \in \mathbb{G}$: remplacer la colonne de X_j par un vecteur ne contenant que des v. Puis calculer les prédictions du modèle pour chaque ligne. Enfin, faire la moyenne de toutes les prédictions pour un v donné. Ainsi à chaque $v \in \mathbb{G}$ est associé une moyenne de prédiction : $\mathbb{E}_j$.
- Tracer la courbe PDP (les v en abscisse et les $\mathbb{E}_j$ en ordonnée).

Ainsi, le but est de mesurer l'effet moyen de la variation de X_j sur les prédictions. Néanmoins, le fait de remplacer les valeurs de v pour toutes les instances fait apparaître le problème d'instances irréalistes. Par exemple, si X_j est l'année du permis en considérant $v = 1980$, remplacer cette valeur de v pour des instances/individus de 25 ans n'a aucun sens. Ici la corrélation entre l'année du permis et l'âge pose un problème de cohérence pour cette méthode.

Une solution serait de construire les M-plot. Dans ce cadre, la valeur moyenne des prédictions pour un v donné est calculé en moyennant les prédictions sur des instances ayant des valeurs de X_j proches de v. De la sorte, la distribution conditionnelle est utilisée et les instances irréalistes sont exclues. Toutefois, cette procédure prend toujours en compte les interactions entre les variables. Par exemple, supposons que l'année du permis n'a aucun effet sur le risque d'accident en auto mais que l'âge a, lui, un effet. Pour les M-plots, la variable année du permis aura toujours un effet du fait de sa corrélation avec la variable âge.

L'approche ALE reprend celle du M-plot en calculant pour sa part une différence entre les prédictions en lieu et place des moyennes.

En partant de la formule des PDP introduite par Greenwell[8] pour x_j de la distribution marginale de X_j :

$$\widehat{f}_{X_j,PDP_{x_j}} = \mathbb{E}[\widehat{f}(X_j, X_{\setminus j})] = \int_{x_{\setminus j} \in X_{\setminus j}} \widehat{f}(x_j, x_{\setminus j}) \mathbb{P}_{X_{\setminus j}}(x_{\setminus j}) dx_{\setminus j}$$

Où $\setminus j$ est l'ensemble des variables privé de la variable X_j. $X_{\setminus j}$ est donc considéré comme étant une variable aléatoire.

Dans le cas du M-plot, la distribution conditionnelle est utilisée :

$$\widehat{f}_{X_j,M_{xj}} = \mathbb{E}_{X_{\setminus j}|X_j}[\widehat{f}(X_j, X_{\setminus j})|X_j = x_j] = \int_{x_{\setminus j} \in X_{\setminus j}} \widehat{f}(x_j, x_{\setminus j}) \mathbb{P}_{X_{\setminus j}}(x_{\setminus j}|X_j = x_j) dx_{\setminus j}$$

Cette formule fournit les moyennes sur les distributions conditionnelles. L'étape finale est d'introduire la moyenne des variations à la place des moyennes. Pour cela, le gradient $\widehat{f}_{(j)}(x_j, x_{\setminus j}) = \frac{\partial \widehat{f}(x_j, x_{\setminus j})}{\partial x_j}$ est introduit. Il permet de mesurer la fluctuation des prédictions. L'estimateur ALE s'écrit :

$$\begin{aligned} \widehat{f}_{X_j, ALE_{x_j}} &= \int_{z_{0,j}}^{z_{N_j,j}} \mathbb{E}_{X_{\setminus j}|X_j}[\widehat{f}_{(j)}(X_j, X_{\setminus j})|X_j = x_j] \\ &= \int_{z_{0,j}}^{z_{N_j,j}} \int_{x_{\setminus j} \in X_{\setminus j}} \{\widehat{f}_{(j)}(X_j, X_{\setminus j}) \mathbb{P}_{X_{\setminus j}}(x_{\setminus j}|X_j = x_j) dx_{\setminus j}\} dz_j \end{aligned} \tag{A.1}$$

Avec $\mathbb{G} = \{z_{0,j}, \ldots, z_{1,j}, \ldots, z_{N_j-1,j}, \ldots, z_{N_j,j}\}$, la grille des valeurs de X_j formée et $N_j + 1$ le nombre de découpages effectués dans la grille. La construction de cette grille peut se faire à l'aide des quantiles.

L'utilisation des dérivations permet d'isoler l'effet de la variable de celle des interactions. L'intégrale supplémentaire permet d'additionner les effets pour tout x_j de X_j. En général, l'ALE est centré à l'aide d'une constante. En pratique, les modèles ne possèdent pas tous des gradients. L'estimation est effectuée en utilisant des intervalles de valeurs. Pour une variable explicative quantitative, l'effet non centré s'estime de la façon suivante :

$$\widehat{f}_{j,ALE}(x) = \sum_{k=1}^{N_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in \mathbb{V}_j(k)} \{f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)})\}$$

La somme sur $\mathbb{V}_j(k)$ permet pour un intervalle donné, de calculer la moyenne des différences de prédictions. Cette moyenne est obtenue en remplaçant la variable X_j par les éléments z de la grille. La somme de ces différences est ensuite divisée par le nombre d'instances pour obtenir la moyenne. $\mathbb{V}_j(k)$ est considéré comme étant un voisinage représentant les instances réalistes avec celle de la valeur de la grille (des valeurs proches de l'instance). Cette somme représente la prise en compte des effets locaux. La première somme permet d'effectuer une sommation sur l'ensemble des intervalles. Ceci est l'effet cumulé.

L'effet obtenu est ensuite centré de la manière suivante :

$$\widehat{f}_{j,ALE}(x) = \widehat{f}_{j,ALE}(x) - \frac{1}{n} \sum_{i=1}^{n} \widehat{f}_{j,ALE}(x^{(i)})$$

Cela permet d'obtenir des effets interprétables par rapport à une valeur moyenne des prédictions. Ainsi, un ALE de 10 dans le cas où $x_j = 5$ signifie que pour cette valeur de x_j, la prédiction est supérieure de 10 à la prédiction moyenne.

Dans le cas où la variable est qualitative ordinale, l'ALE se calcule de la même manière que dans le cas quantitatif. En effet, le caractère ordinal permet de cumuler les effets dans un certain ordre, de retrouver un découpage en grille de la variable X_j.

Pour le cas où les variables ne sont pas nativement ordonnées, il est impératif d'avoir recours à des tables de fréquences relatives ou de distances comme celle de Kolmogorov-Smirnov dans le but de calculer un certain ordre entre les variables.

En plus de prendre en compte les corrélations et d'isoler l'effet des interactions, les ALE se calculent plus rapidement que les PDP. Ils permettent aussi de visualiser les effets de l'interaction entre deux variables sur le modèle en construisant un ALE 2D.
Toutefois, cette méthode introduit son lot d'inconvénients. D'abord, l'interprétation d'un intervalle à un autre de la grille n'est pas intuitive. En effet, les effets ne sont interprétables que localement (c'est à dire intervalle de la grille par intervalle). Ils sont simplement cumulés pour l'obtention d'une courbe.
De plus si les variables, en plus d'être corrélée, interagissent, alors l'attribution des effets d'interactions n'est pas très clair. En effet dans "Model-Agnostic Effects Plots for Interpreting Machine Learning Models", il est prouvé que l'attribution par la méthode ALE des contributions des variables dans un modèle linéaire est différente des contributions intuitives.
Ensuite, le choix du nombre d'intervalle peut s'avérer difficile. Plus, il y a d'intervalles plus l'ALE est instable et plus l'ALE laisse passer les subtilités contenues dans les données.
Enfin, il n'existe encore aucune méthode complémentaire à l'ALE permettant d'obtenir les contributions locales comme dans le cas des PDP (voir ICE). Toutefois, les améliorations apportées par cette méthode dans le traitement des corrélations et des interactions placent cette méthode au dessus de la méthode PDP dans quasiment tous les cas d'application.

A.3 Méthodes d'interprétabilité agnostique locale

Les méthodes d'interprétabilité locales permettent d'expliquer la part jouée par chaque variable du modèle dans l'obtention d'une prédiction précise.

A.3.1 Individual Conditional Expectation ICE

Les ICE représentent les briques qui permettent de construire les PDP. En effet, dans un graphique ICE, chaque ligne représente l'effet de la modification de la variable à expliquer sur les prédictions pour une instance. Ainsi, en lieu et place de la moyenne réalisée dans le cas des PDP, toutes les instances sont représentées.
Ainsi, $\forall i \in \{1, ..., n\}, \widehat{f}(x_j, x_{\setminus j}^{(i)})$ est calculé avec x_j qui varie sur un ensemble des valeurs de X_j et $x_{\setminus j}^{(i)}$ qui reste fixé[30].

Cette approche est encore plus intuitive que celle des PDP. Une ligne représente une courbe d'espérance conditionnelle individuelle, soit l'évolution des prédictions pour un individu donné quand les modalités de la variable explicative change.
Néanmoins, l'utilisation de cette méthode pour plus d'une variable à la fois ne présente aucun intérêt car les représentations ne sont pas exploitables. Il n'est donc pas possible d'analyser les effets croisés entre les variables. Et même dans le cas à une variable,

tracer un trop grand nombre de courbes rendrait le graphique inexploitable. De plus, des instances irréalistes peuvent apparaître dans le cas où les variables sont corrélées.

A.3.2 LIME

La méthode LIME a été introduite en 2016 par Ribeiro et al.[54]. L'idée sous jacente de cette méthode est de construire au niveau local des modèles de substitution plus simple qui permettent d'approcher le modèle initial dans un certain voisinage. Ainsi, pour une instance donnée et la prédiction associée, la méthode LIME a pour but de construire une fonction g qui permettra d'approcher le comportement du modèle au voisinage de cette instance. Pour cela :

- Les variables explicatives de l' instance sont permutées plusieurs fois.
- Les prédictions sont recalculées pour chaque nouvelle instance obtenue après permutation.
- Un score de similarité est calculé entre les permutations prises une à une et l'instance initiale. Ce score est calculé en utilisant les distances appropriées.
- Un sous ensemble de variables explicatives significatives pour la modélisation des prédictions sur les données permutées est sélectionné.
- Le modèle de substitution est choisi et calibré. Ce modèle est généralement un modèle linéaire multiple ou pénalisé (LASSO, RIDGE).
- Les coefficients du modèle simplifié sont extraits et utilisés comme explication.

L'utilisation de modèles de substitution simples et parcimonieux permet d'obtenir des explications rapides et simples, tout en gardant en vue la qualité de l'approximation fournie (en regardant la précision du modèle local). Toutefois, la définition du voisinage d'une instance est assez subjective. De plus, les instances obtenues après perturbations ne sont pas toujours cohérentes car elles ne prennent pas en compte les corrélations qui existent entre les variables. Les explications obtenues sont aussi assez instables du fait de l'échantillonnage effectué dans la construction des instances perturbées. Il a aussi été prouvé[49] que les explications du modèle LIME pouvait être intentionnellement manipulées.

A.3.3 Valeur shapley et SHAP

SHAP (SHapley exPlainable Value) est une méthode qui utilise les valeurs shapley introduites en théorie des jeux par Shapley en 1953[72].
Soit un jeu coopératif ou collaboratif où p joueurs collaborent dans le but d'obtenir un gain. Survient alors le problème de l'attribution équitable du gain aux p joueurs du jeu. En d'autres termes quel est l'apport de chaque joueur au gain obtenu. C'est dans le but de répondre à cette problématique que les valeurs shapley ont été introduites. Dans la suite, le terme contribution désignera la part de chaque individu dans le gain obtenu. Il se notera $\phi_j(v)$ pour désigner la contribution du joueur j au gain v. L'attribution désignera l'action de calculer les contributions de chaque joueur du jeu. La notion de répartition équitable a été étudié et 4 points ont été retenus pour définir un paiement équitable :

— **Efficacité** : La somme des contributions des variables doit être égale à la différence entre la prédiction moyenne et la prédiction pour l'instance concernée.

$$\sum_{j=1}^{p} \phi_j(v) = \widehat{f}(x) - \mathbb{E}_X(\widehat{f}(X)).$$

— **Symétrie** : Deux variables contribuant de la même manière dans toutes les combinaisons (coalitions) ont toujours la même valeur shapley. Si

$$\forall S \subseteq \{x_1, ..., x_p\} \backslash \{x_j, x_k\},\ v(S \cup x_k) = v(S \cup x_j),$$

alors :

$$\phi_j = \phi_k.$$

— **Facticité** : Une variable n'ayant aucun effet sur les valeurs prédites pour toutes les combinaisons de valeurs à laquelle elle est ajoutée doit avoir une valeur Shapley de 0.

$$v(S \cup x_j) = v(S), \quad \forall S \subseteq x_1, ..., x_p \backslash \{x_j\} \implies \phi_j = 0.$$

— **Additivité** : Dans un jeu où les gains sont somme de gains, les valeurs shapley sont la somme des valeurs shapley. Par exemple, dans le cas d'un modèle ensembliste où la prédiction est la moyenne des prédictions de plusieurs modèles, la valeur shapley est aussi la moyenne des valeurs dans chaque modèle.

Le parallèle est fait entre le cadre théorique initial de la théorie des jeux et celui de l'interprétabilité des modèles. Le jeu est la prédiction d'une certaine variable cible à l'aide d'un vecteur de valeurs (une instance d'un jeu de données), les joueurs sont les valeurs des variables explicatives et le gain est la différence entre la prédiction et la valeur moyenne de la variable cible dans le jeu de données.

Ainsi, la valeur shapley de la variable j dans l'obtention du gain g se définit comme étant sa contribution (au gain total), additionnée et pondérée sur toutes les combinaisons possibles des valeurs des variables :

$$\phi_j(v) = \sum_{S \subseteq x_1, ..., x_p} \frac{|S-1|!(p-|S|)!}{p!}(v(S) - v(S \backslash \{x_j\}))$$

où S est un sous-ensemble des réalisations des variables explicatives, $|S|$ le cardinal de S.

Dans la pratique, l'évaluation de toutes les combinaisons de valeurs de variables avec et sans la j^{eme} observation demande des temps de calculs exponentiels. En 2014, Strumbelj[85], proposent l'utilisation d'un échantillonage par Monte Carlo comme approximation de la valeur shapley exacte :

$$\widehat{\phi_j} = \frac{1}{M} \sum_{m=1}^{M} (\widehat{f}(x_{+j}^m) - \widehat{f}(x_{-j}^m)),$$

avec j l'indice de la variable à expliquer et M le nombre d'itérations. x^m_{+j} est une instance dans laquelle les valeurs d'un nombre aléatoire de variables ont été remplacés par des points choisis aléatoirement. Toutes les variables peuvent être choisies sauf la variable j. Pour x^m_{-j}, la valeur de la variable j est aussi tirée dans un échantillon.
La valeur shapley est plébiscitée car elle s'appuie sur des fondements théoriques solides et prouvés. Les propriétés de répartitions équitables du paiement assurent la qualité des explications fournies. Il se pourrait qu'elle soit une des seules méthodes d'interprétabilité assez solide pour être admise du point de vue réglementaire.
Cette méthode permet aussi une explication par rapport à n'importe quel instance du jeu de données. En effet, une prédiction peut être expliquée relativement à la valeur moyenne, ou un point spécifique ou un sous-ensemble du jeu de données.

Malgré tout, la complétude de cette méthode se paye au prix d'un temps de calcul exponentiel. L'utilisation de l'approximation de Monte Carlo permet d'obtenir des temps plus réalistes mais toujours élevés. De plus, l'approximation introduit de la variance dans l'estimation des valeurs shapley. D'autre part, cette méthode aussi ne permet pas d'analyser l'effet du changement de la valeur des variables sur la prédiction.
Enfin, l'utilisation des permutations introduit le problème des instances irréalistes dans le cas où les variables sont corrélées.

La méthode de SHAP est une extension des valeurs shapley dans le cadre de l'interprétation des prédictions des modèles d'apprentissage automatique. Dans cette méthode, les contributions de chaque variable sont représentées sous la forme d'un modèle linéaire. Ainsi les explications se présentent de la façon suivante :

$$g(z^{'}) = \phi_0 + \sum_{j=1}^{M} \phi_j z^{'}_j,$$

où g le modèle d'explication, $z^{'} \in \{0,1\}^M$ le vecteur indiquant la présence d'une variable dans la combinaison, M est la taille de la combinaison et ϕ_j la contribution de la valeur de la variable X_j. Cette forme permet de conserver toutes les propriétés des valeurs shapley et respectant 3 propriétés supplémentaires.
La première est la propriété dite de **précision locale**. Ainsi,

$$f(x) = g(x^{'}) = \phi_0 + \sum_{j=1}^{M} \phi_j x^{'}_j.$$

Une prédiction peut se décomposer comme la somme des contributions individuelles. Pour $\phi_0 = \mathbb{E}_X(\widehat{f}(X))$ et $x_j = 1\ \forall j$, cette propriété redonne celle d'efficacité des valeurs shapley.

Les termes manquants spécifie que lorsqu'une variable est absente d'une combinaison, elle obtient une contribution de 0.
La cohérence est une propriété qui assure que si un modèle change de manière à ce que la contribution marginale d'une valeur de variable augmente (resp. reste la même) indé-

pendamment des autres variables, la valeur shapley de cette variable augmente également (resp. reste la même).

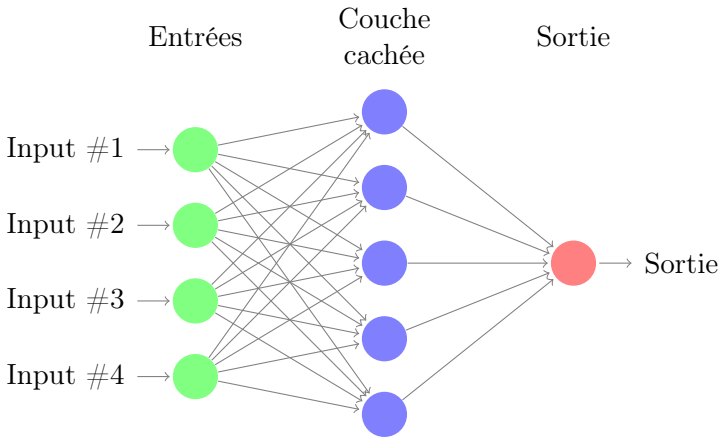
En plus de conserver tous les avantages des valeurs shapley, la méthode SHAP, en utilisant la décomposition additive des explications, permet d'obtenir l'effet global des variables sur le modèle. En effet, en traçant pour une variable explicative fixée, les valeurs shapley de tous les individus, il est possible d'obtenir de la même manière qu'avec les courbes ICE l'effet global de la variable. La méthode SHAP permet donc d'unifier les valeurs shapley et les autres méthodes d'explicabilité locales (LIME, ICE). De plus, la méthode SHAP possède une variante pour les modèles à arbre : la méthode TreeSHAP. Cette variante permet d'obtenir des temps de calculs plus raisonnables pour les modèles à arbres. Cela rend plus accessible les calculs requis pour une interprétation globale du modèle (importance des variables, interactions, dépendance etc.). Ces interprétations étant basées sur les valeurs shapley (qui permettent les interprétations locales) une cohérence est conservée. Ce n'est pas le cas pour des méthodes comme LIME.

Toutefois, cette variante peut produire des explications non intuitives. En effet, dans le but d'accélérer les calculs, la méthode TreeSHAP modifie légèrement le jeu ce qui peut conduire à des attributions de valeur TreeShap non nulles pour des variables non influentes. Comme les valeurs shapley, cette méthode demande des temps de calculs élevés et ne permet pas de mesurer l'effet du changement d'une variable sur les prédictions. Il a aussi été prouvé[22] que la méthode SHAP peut être utilisée dans le but de produire intentionnellement des interprétations trompeuses.

A.4 Interprétabilité propres aux réseaux de neurones

Les réseaux de neurones artificiels sont des modèles d'apprentissage automatiques dont la forme est inspirée par les réseaux neuronaux des cerveaux de mammifères. Il s'agit d'un ensemble de neurones inter-connectés. Ces neurones reçoivent des valeurs réelles provenant des neurones auxquels ils sont connectés. Ils traitent ces valeurs à l'aide d'une fonction. Des poids sont associés à ces neurones et à leurs sorties. Tout au long du processus d'entraînement, un gradient est utilisé pour modifier la valeur des poids de chaque neurone. Ainsi, le modèle "apprend" en modifiant ces poids dans le but de réduire l'erreur à la sortie du modèle. L'entraînement de ces modèles se fait en général par la méthode de Rétropropagation du gradient (backpropagation). Cette méthode permet un calcul efficient du gradient par rapport à tous les poids du réseau en utilisant le théorème de dérivation des fonctions composées. Une fois le gradient calculé avec cette méthode, des méthodes comme SGD (descente du gradient stochastic) ou Adam permettent de mettre à jour les poids.

Ces modèles sont appelés modèles d'apprentissage profonds (deep learning) du fait du grand nombre de couches cachées. Ces structures permettent d'extraire les informations des données et ensuite de les utiliser pour la tâche requise. En effet, dans les modèles d'apprentissage statistique classiques, en amont de la modélisation, l'étape d'extraction des caractéristiques (feature engineering) est indispensable pour transformer les données brutes en données de prédiction. Les données brutes doivent être transformées en carac-



téristiques représentant plus précisément le phénomène à modéliser. Dans les modèles de deep learning, ce travail est directement effectué dans les couches cachées. En effet, par leur structure, ces modèle sont capables de représenter les variables dans de très grandes dimensions et ainsi, mettre en forme les données dans le but d'en tirer le maximum d'informations. Ces nouvelles représentations de l'information s'effectuent donc à l'intérieur du modèle sans aucune transparence et sont donc inexploitables pour expliquer le comportement du modèle. Pour interpréter ces modèles, il est possible de se tourner vers les méthodes agnostiques présentées dans la section précédente. Néanmoins, l'existence du gradient peut permettre l'implémentation de méthodes plus efficiente en terme de temps de calcul. De plus, dans ces modèles, les couches cachées et neurones apprennent des caractéristiques particulières des données qu'il serait intéressant de mettre en lumière.
En plus d'utiliser le gradient pour améliorer les approches d'interprétabilité classique, les réseaux peuvent être construit dans le but d'être directement interprétables.
La littérature est assez riche dans le cas des images et du texte mais un peu moins pour les données tabulaire.

A.4.1 Méthodes d'interprétation basées sur les gradients

Dans les sections précédentes, la méthode ALE a été présentée. Sous sa forme continue, elle est calculable en utilisant des gradients. En 2016, Shrikumar et al.[73] introduisent pour la première fois l'utilisation d'attributions marginales en considérant les dérivées directionnelles :

$$x_j \widehat{f}_j(x) = x_j \frac{\partial \widehat{f}(x)}{\partial x_j}.$$

D'un point de vue mathématiques, ce calcul des attributions marginales peut s'apparenter à une décomposition additive des contributions à travers des décompositions de Taylor. En considérant par exemple un modèle linéaire de coefficients $\{\beta_0, \ldots, \beta_n\}$, ces coefficients s'assimilent à une mesure de l'importance des variables associées. Comme discuté dans les sections précédentes, une telle décomposition est celle qui serait idéale pour le cadre assurantiel traité dans ces travaux. Néanmoins, en dehors des modèles à forme

linéaire, de telles décompositions ne sont possibles que localement (ALE par exemple). L'utilisation des valeurs shapley permet de prendre en compte les non linéarités et les interactions entre les variables.
Des méthodes tels que LRP (Layer-wise propagation)[5] and Deep LIFT (Deep Learning Important FeaTures)[74] ont été développées spécialement pour les modèles de deep learning, l'idée étant qu'en partant des prédictions et en remontant le réseau, une certaine importance peut être attribuée à chaque entrée. Pour cela, un indice d'importance est redistribué récursivement de couche en couche. En 2019, Ancona et al.[14] discutent du fait que ces méthodes s'apparentent à une moyenne des attributions marginales.

Des méthodes comme Integrated Gradient[14] permettent aussi d'identifier les entrées les plus importantes pour une prédiction en utilisant le gradient du modèle. En partant d'une instance de référence, une interpolation est réalisée entre l'instance de référence et l'instance à expliquer. Ces interpolations sont de petites variations des variables explicatives. Le gradient est ensuite utilisé pour mesurer l'effet de ces changements sur les prédictions. L'importance des variables est obtenue en moyennant les valeurs des gradients sur l'ensemble d'interpolation.
En théorie ces méthodes peuvent être utilisées pour tout type de problèmes (image, texte, tableau). Elles sont toutefois plus adaptées aux images.

Des méthodes comme Marginal Attribution by Conditioning on Quantiles (MACQ)[58] ont été construites pour une utilisation sur les données tabulaires. Ils utilisent les attributions marginales qu'ils "agrègent" dans le but d'obtenir des importances globales, le but étant de fixer des seuils de prédictions et de décrire l'effet des variables sur ces seuils sans perturber les instances (sans faire de permutations).

A.4.2 Structures interprétables par construction

Au lieu d'interpréter un modèle boite noire après sa construction, il est possible de construire des structures plus interprétables. C'est dans ce sens que He et al.[39] introduisent en 2016 le modèle ResNet construit autour d'un terme linéaire.
Il existe aussi le modèle LassoNet[69] qui effectue une pénalisation Lasso sur les éléments du réseau de neurones. Ces deux réseaux ont la particularité d'être construits dans le but de permettre à certaines informations de sauter certaines couches cachées (de ne pas être traitées par celle ci).

Des approches comme celle du GLMnet[68] utilisent cette construction autour d'un terme linéaire en y ajoutant des couches permettant la modélisation de toute forme d'interactions et de comportements non linéaires. D'autres méthodes utilisées dans xNN ou NAM restreignent les RN en modifiant la manière dont les caractéristiques extraites des données circulent dans le réseau[28, 77]. Pour terminer cette liste non exhaustive, il est possible de citer les CAXNN[67] qui sont une extension des xNN. Leur objectif est d'inclure des composantes linéaires dans un modèle combiné.

Bibliographie

[1] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. *A reductions approach to fair classification.* Icml'18 (pp. 60–69), 2018.

[2] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. *Fair Regression : Quantitative Definitions and Reduction-based Algorithms.* Proceedings of the 36 th International Conference on Machine Learning, 2019.

[3] M. Agueh and G. Carlier. *Barycenters in the wasserstein space. SIAM Journal on Mathematical Analysis.* SIAM Journal on Mathematical Analysis, 43(2) :904–924, 2011.

[4] Leigh Alexander. *Do Google's 'unprofessional hair' results show it is racist ?* The Guardian, 2016.

[5] A. Binder, S. Bach, G. Montavon, Müller K.-R., and W. Samek. *Layer-wise relevance propagation for deep neural network architectures.* Information Science and Applications (ICISA). Kim K., Joukov N. (Eds.). Springer, Lecture Notes in Electrical Engineering 376, 2016.

[6] R. Binns. *The apparent conflict between individual and group fairness.* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 514–524, 2020.

[7] E. Black, S. Yeom, and et M. Fredrikson. *FlipTest : Fairness Testing via Optimal Transport.* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAccT '20, Barcelona, Spain, pp. 111–121, 2020.

[8] Greenwell Brandon, Bradley Boehmke, and Andrew McCarthy. *A simple and effective modelbased variable importance measure.* arXiv :1805.04755, 2018.

[9] L. Breiman and Friedman. *Estimating optimal transformations for multiple regression and correlation.* Journal of The American Statistical Association, 80 :580–598, 09 1985a, 1985.

[10] Leo Breiman. *Random Forest.* Kluwer Academic Publishers, 2001.

[11] Flavio Calmon, Dennis Wei, and Bhanukiran Vinzamuri. *Optimized Pre-Processing for Discrimination Prevention.* Advances in Neural Information Processing Systems 30, 2017.

[12] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, and Daniele Regoli. *A clarification of the nuances in the fairness metrics landscape.* Scienctific Reports, Data science and AI, 2022.

[13] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth Vishnoi. *Classification with Fairness Constraints : A Meta-Algorithm with Provable Guarantees.* arXiv, 2018.

[14] Ancona Ceolini and Öztireli Gross. *Gradient-based attribution methods.* Explainable AI : Interpreting, Explaining and Visualizing Deep Learning. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., M¨uller K.-R. (Eds.). Springer, Lecture Notes in Artificial Intelligence 11700, 168-191, 2019.

[15] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. *Bias in Machine Learning Software : Why ? How ? What to Do ?* arXiv, 2021.

[16] A. Chouldechova. *Fair prediction with disparate impact : A study of bias in recidivism prediction instruments.* Big data, 5(2) :153–163, 2017.

[17] André Comte-Sponville. *Dictionnaire philosophique.* Presses Universitaires de France, 2013.

[18] Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, and al. *Fairness Measures for Machine Learning in Finance.* arXiv, 2020.

[19] Conseil de l'Union Européenne. *Proposition de directive du Conseil relative à la mise en œuvre du principe de l'égalité de traitement entre les personnes sans distinction de religion ou de convictions, de handicap, d'âge ou d'orientation sexuelle.* Presses Universitaires de France, 2008.

[20] Laurent Dupont, Olivier Fliche, and Su Yang. *Gouvernance des algorithmes d'intelligence artificielle dans le secteur financier.* Pôle Fintech-Innovation, ACPR, 2020.

[21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. *Fairness Through Awareness.* arXiv :1104.3913, 2011.

[22] Slack Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. *Fooling lime and shap : Adversarial attacks on post hoc explanation methods.* In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180-186, 2020.

[23] Cook et al. *The Gender Earnings Gap in the Gig Economy : Evidence from over a Million Rideshare Drivers.* National Bureau of Economic Research, 2018.

[24] Michael Feldman, Sorelle Friedler, and John Moeller. *Certifying and Removing Disparate Impact.* Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 259–268, 2015.

[25] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. *All Models are Wrong, but Many are Useful : Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously.* arXiv :1801.01489, 2018.

[26] Samsung for business. *Your phone is now more powerful than your pc.* Samsung Insights, https ://insights.samsung.com/2021/08/19/your-phone-is-now-more-powerful-than-your-pc-3/, 2021.

[27] Y. Freund and R. Schapire. *Game theory, on-line prediction and boosting.* Proceedings of the Ninth Annual Conference on Computational Learning Theory, 1996.

[28] Agarwal Frosst, Zhang Caruana, and Hinton G. *Neural additive models : interpretable machine learning with neural nets.* arXiv, 2020.

[29] S. Galhotra, Y. Brun, and A. Meliou. *Fairness Testing : Testing Software for Discrimination.* arXiv : 1709.03221, 2017.

[30] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. *Peeking Inside the Black Box : Visualizing Statistical Learning with Plots of Individual Conditional Expectation.* Journal of Computational and Graphical Statistics, 2017.

[31] P. Gordaliza, E. Del Barrio, F. Gamboa, and et J. Loubes. *Obtaining fairness using optimal transport theory.* International Conference on Machine Learning, pages 2357–2365, 2019.

[32] T. Le Gouic and J-M. Loubes. *Fair regression in l2 : optimal fair prediction for demographic parity.* arXiv, 2020.

[33] V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki. *Fairness-aware neural rényi minimization for continuous features.* Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI 2020 (pp. 2262–2268), 2020.

[34] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. *Fairness without the sensitive attribute via Causal Variational Autoencoder.* arXiv, 2021.

[35] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. *Kernel methods for measuring independence.* J. Mach. Learn. Res., 6 :2075–2129, 2005.

[36] Ulrike Grömping. *Model-Agnostic Effects Plots for Interpreting Machine Learning Models.* Reports in Mathematics, Physics and Chemistry : Department II, Beuth University of Applied Sciences Berlin, 2020.

[37] David R Hardoon and John Shawe-Taylor. *Convergence analysis of kernel canonical correlation analysis : theory and practice.* Machine learning, 74(1) :23–38, 2009.

[38] Moritz Hardt, Eric Price, and Nathan Srebro. *Equality of Opportunity in Supervised Learning.* arXiv :1610.02413v1, 2016.

[39] K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognitions.* 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[40] C. Ilvento. *Metric learning for individual fairness.* arXiv 1906.00250, 2019.

[41] Phoebe Jackson-Edwards. *Student is left shocked after a Google search for 'unprofessional hairstyles at work' features ALL black women - while 'professional styles' lists just white models.* Daily Mail UK, 2016.

[42] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. *Wasserstein fair classification.* Thirty-Fifth Uncertainty in Artificial Intelligence Conference, 2019.

[43] C. Jung and al. *Eliciting and enforcing subjective individual fairness.* arXiv 1905.10660, 2019.

[44] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. *Decision Theory for Discrimination-Aware Classification.* 2012 IEEE 12th International Conference on Data Mining, 2016.

[45] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. *Preventing Fairness Gerrymandering : Auditing and Learning for Subgroup Fairness.* Lecture Notes in Computer Science book series (LNAI,volume 7524), 2012.

[46] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. *Preventing Fairness Gerrymandering : Auditing and Learning for Subgroup Fairness.* arXiv, 2018.

[47] Denis Kessler, Amélie de Montchatlin, and Christian Thimann. *Assurance et développement économique : croissance, stabilisation et répartition.* Impact Insurance, publication N°46, 2016.

[48] S. Lapuschkin, S. Waldchen, A. Binder, G. Montavon, Samek W., and R. Müller. *Unmasking clever hans predictors and assessing what machines really learn.* Nature Communications 10, 1096, 2019.

[49] T. Laugel and X. Renard. *Defining Locality for Surrogates in Post-hoc Interpretablity.* arXiv :1806.07498, 2018.

[50] W.N. Locke and D.A. Booth. *Machine Translation of Languages.* Cambridge (Massachusetts), MIT Press, 1955, 15–23 p., 1955.

[51] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. *The Randomized Dependence Coefficient.* arXiv, 2013.

[52] Ribeiro M., Singh S., and Guestrin C. *Why should i trust you ? : Explaining the predictions of any classifier.* ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144, 2016.

[53] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. *Survey on Causal-based Machine Learning Fairness Notions.* arXiv, 2020.

[54] Ribeiro Marco, Sameer Singh, and Carlos Guestrin. *Why should I trust you ? : Explaining the predictions of any classifier.* Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016.

[55] J. Mary, C. Calauzènes, and Karoui. *Fairness-aware learning for continuous attributes and treatments.* ICML'19, 4382–4391, 2019.

[56] D. McNamara, C. S. Ong, and R. C. Williamson. *Provably fair representations.* arXiv 1710.04394, 2017.

[57] Ninareh Mehrabi and Fred Morstatter. *Survey on Bias and Fairness in Machine Learning.* arXiv, 2022.

[58] Michael Merz, Ronald Richman, Andreas Tsanakas, and Mario Wüthrich. *Interpreting Deep Learning Models with Marginal Attribution by Conditioning on Quantiles.* arXiv, 2021.

[59] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. *Algorithmic fairness : Choices, assumptions, and definitions.* Annual Review of Statistics and Its Application 8 (2021) : 141-163, 2021.

[60] Christoph Molnar. *Interpretable Machine Learning, 3-8.* 2 edition, 2022.

[61] A. Mordvintsev, C. Olah, and M. Tyka. *Inceptionism : Going deeper into neural networks.* 2015.

[62] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. *Social data : Biases, methodological pitfalls, and ethical boundaries.* arXiv, 2016.

[63] J. Pearl, M. Glymour, and N. Jewell. *Causal Inference in Statistics : A Primer.* Wiley, 2016.

[64] J. Pearl and D. Mackenzie. *The Book of Why : The New Science of Cause and Effect.* Basic Books, 2018.

[65] Oskar Pfungst. *Clever Hans (The Horse of Mr. Von Osten).* Projet Gutenberg, 2010.

[66] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Weinberger. *On Fairness and Calibration.* Advances in Neural Information Processing Systems 30 (NIPS), 2017.

[67] R. Richman. *Mind the gap - safely incorporating deep learning models into the actuarial toolkit.* SSRN Manuscript ID 3857693, 2021.

[68] Ronald Richman and Mario Wuthrich. *LocalGLMnet : interpretable deep learning for tabular data.* arXiv, 2021.

[69] Lemhadri Ruan and Abraham Tibshirani. *LassoNet : a neural network with feature sparsity.* Journal of Machine Learning Research 22, 1-29, 2021.

[70] A Rényi. *On measures of dependence.* Acta mathematica hungarica 10(3-4) :441–451, 1959.

[71] Wojciech Samek and Klaus-Robert Müller. *Towards Explainable Artificial Intelligence.* arXiv, 2019.

[72] Lloyd S. Shapley. *A value for n-person games.* Contributions to the Theory of Games 2.28, 1953.

[73] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. *Not just a black box : learning important features through propagating activation différences.* arXiv, 2016.

[74] M. Sundararajan, A. Taly, and Q. Yan. *Axiomatic attribution for deep networks.* Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, PMLR. International Convention Centre, Sydney, Australia, 70, 3319-3328, 2017.

[75] Harini Suresh and John Guttag. *A Framework for Understanding Unintended Consequences of Machine Learning.* arXiv :1901.10002, 2019.

[76] Gabor J Székely, Maria L Rizzo, and al. *Brownian distance covariance.* The annals of applied statistics, 3(4) :1236–1265, 2009.

[77] J. Vaughan, A. Sudjianto, E. Brahimi, J. Chen, and V. Nair. *Explainable neural networks based on additive index models.* arXiv, 2018.

[78] S. Wachter, B. Mittelstadt, and C. Russell. *Why Fairness Cannot Be Automated : Bridging the Gap Between EU Non-Discrimination Law and AI.* SSRN Scholarly Paper ID 3547922, Social Science Research Network, 2020.

[79] Wikipedia. *Bias–variance tradeoff.* https ://en.wikipedia.org/wiki/Bias-variance$_t radeoff$, 2013.

[80] H. S. Witsenhausen. *On sequences of pairs of dependent random variables.* SIAM J. Appl. Math., 28 :100–113, 1975.

[81] B. Woodworth, S. Gunasekar, M. Ohannessian, and N. Srebro. *Learning nondiscriminatory predictors.* arXiv preprint arXiv :1702.06081, 2017.

[82] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. *PC-Fairness : A Unified Framework for Measuring Causality-based Fairness.* arXiv, 2019.

[83] M. Zafar, I. Valera, M. Gomez Rodriguez, and K. Gummadi. *Fairness beyond disparate treatment disparate impact : Learning classification without disparate mistreatment.* Proceedings of the 26th International Conference on World Wide Web, pages 1171–1180, 2017.

[84] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C Dwork. *Learning fair representations.* International Conference on Machine Learning, 325–333, 2013.

[85] Erik Štrumbelj and Igor Kononenko. *Explaining prediction models and individual predictions with feature contributions.* 2014.