Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 05/03/2025

Par : **Marc Jodel SIMO**

Titre : **Equité des modèles en assurance, mesure et mitigation des discriminations.**

Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)

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

Membres présents du jury de la filière

Nom :

Membres présents du jury de l'Institut des Actuaires

Entreprise : BNP Paribas Cardif

Signature :

Directeur du mémoire en entreprise :

Nom :
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

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Remerciements

Ce travail est bien plus qu'un effort individuel ; il résulte de nombreuses contributions précieuses, souvent discrètes mais essentielles. À ce titre, je souhaite exprimer ma profonde gratitude à toutes les personnes qui ont, de près ou de loin, apporté leur soutien et leur expertise à la réalisation de ce mémoire.

Tout d'abord, je remercie BNP Paribas Cardif de m'avoir offert l'opportunité d'effectuer ce travail dans un cadre aussi enrichissant. Je tiens particulièrement à exprimer ma reconnaissance à mon encadrant professionnel, Boris Noumedem, Actuaire Data Scientist, pour son accompagnement attentif, ses conseils avisés et la qualité de ses retours tout au long de cette étude. Mes sincères remerciements vont également à Quang Do, Manager Actuaire de l'équipe Predictive Analytics, pour ses orientations précieuses et son regard éclairé sur les problématiques abordées.

Je suis également très reconnaissant envers mon tuteur académique, Caroline Hillairet, Enseignant-Chercheur à l'ENSAE-Paris, pour son suivi rigoureux, sa disponibilité et ses recommandations précieuses, qui ont contribué à structurer et approfondir ce travail.

Enfin, je remercie l'ensemble des équipes du DataLab de BNP Paribas Cardif ainsi que mes proches pour leur soutien et leurs encouragements tout au long de cette expérience.

Résumé

L'intelligence artificielle (IA) joue un rôle croissant dans le secteur de l'assurance, offrant des opportunités inédites pour améliorer l'efficacité, la personnalisation et la gestion des risques. Cependant, l'usage des modèles prédictifs soulève des préoccupations majeures liées à l'équité, car ces outils peuvent reproduire ou amplifier des biais présents dans les données, générant ainsi des discriminations indirectes. Ce mémoire explore ces enjeux en s'appuyant sur des analyses théoriques et empiriques.

Les objectifs de ce travail sont triples : premièrement, comprendre les principes d'équité dans un cadre actuariel et algorithmique ; deuxièmement, évaluer les biais à l'aide de métriques adaptées, telles que la part de variance expliquée et la distance de Kolmogorov ; et troisièmement, proposer des méthodes pour réduire les discriminations, en intégrant des approches telles que l'orthogonalisation des variables explicatives et le transport optimal des distributions.

Les résultats montrent que, malgré la suppression des variables sensibles, les biais persistent souvent par le biais de proxys. Les méthodologies proposées permettent de réduire efficacement ces disparités, mais elles posent des défis en termes d'interprétabilité et de performances des modèles. Ce mémoire propose également des pistes pour élargir l'étude à d'autres critères sensibles, tout en soulignant la nécessité d'un cadre équilibré entre équité, efficacité et transparence.

Mots-clés : Équité, intelligence artificielle, biais, proxy, orthogonalisation, transport optimal.

Abstract

Artificial intelligence (AI) is playing an increasingly significant role in the insurance industry, offering unprecedented opportunities to enhance efficiency, personalization, and risk management. However, the use of predictive models raises critical concerns regarding fairness, as these tools can replicate or amplify biases present in the data, thereby generating indirect discrimination. This thesis addresses these issues through both theoretical and empirical analyses.

The objectives of this work are threefold : first, to understand the principles of fairness within an actuarial and algorithmic framework ; second, to assess biases using appropriate metrics, such as explained variance and Kolmogorov distance ; and third, to propose methods to mitigate discrimination by incorporating approaches like the orthogonalization of explanatory variables and the optimal transport of distributions.

The findings demonstrate that, despite the removal of sensitive variables, biases often persist through proxies. The proposed methodologies effectively reduce these disparities but present challenges in terms of model interpretability and performance. This thesis also outlines avenues for extending the study to other sensitive criteria while emphasizing the need for a balanced framework between fairness, efficiency, and transparency.

Keywords : Fairness, artificial intelligence, biases, proxy, orthogonalization, optimal transport.

Note de Synthèse

Contexte et Problématique

L'intelligence artificielle (IA) joue un rôle croissant dans le secteur de l'assurance, notamment dans la tarification et la gestion des sinistres. Bien que prometteuse en termes de gains d'efficacité et de personnalisation, l'utilisation des modèles prédictifs soulève des préoccupations majeures concernant l'équité. Ces modèles peuvent perpétuer ou amplifier des biais présents dans les données, créant des discriminations indirectes. Ce mémoire explore ces enjeux, avec un accent particulier sur l'assurance automobile, et vise à proposer des méthodes pour détecter et corriger ces biais.

L'étude s'inscrit dans un cadre réglementaire européen strict, marqué par des initiatives telles que la **Directive 2000/78/CE** et l'AI Act. Ces réglementations imposent des obligations de non-discrimination, en particulier en matière de genre, et incitent à développer des modèles transparents et équitables.

Objectifs et Méthodologie

Le mémoire aborde trois objectifs principaux :

— **Comprendre les principes d'équité** : Analyser les concepts d'équité à travers la littérature afin d'orienter les concepteurs et utilisateurs de modèles sur les définitions adaptées à leurs situations.

— **Mesurer l'équité des modèles prédictifs** : Évaluer l'impact des données biaisées et des proxys sur les décisions algorithmiques, grâce à des métriques quantitatives telles que la parité statistique et la distance de Kolmogorov.

— **Proposer des méthodes de mitigation** : Explorer des solutions pour atténuer les biais, comme l'orthogonalisation des variables explicatives et le transport optimal des distributions.

La partie conceptuelle s'appuie sur un ensemble d'articles et de livres dédiés à la définition de l'équité. Elle aboutit à la recommandation d'un arbre de décision pour l'orientation sur le choix de définition d'équité en fonction du cas d'usage ([28]). Quant à la partie expérimentale, elle utilise les données de l'assurance responsabilités civile des particuliers en Belgique, proposée par A. Charpentier et C. Dutang ([12]) dans le package R CASdatasets.

Diagnostic des biais : Tests d'hypothèse pour les proxys

Pour identifier les proxys potentiels de la variable sensible, des tests d'hypothèse statistiques sont effectués :

— **Tests de corrélation** : Les corrélations entre les variables explicatives et la variable sensible (`DRIVER_GENDER`) sont évaluées. Ces analyses permettent de détecter les relations linéaires susceptibles d'introduire des biais.

— **Tests d'indépendance** : Des tests comme le **test de Kolmogorov-Smirnov** ou le **test de Chi-deux** sont utilisés pour vérifier si les distributions des variables explicatives diffèrent significativement en fonction du genre.

— **Analyse de variance (anova)** : Pour les variables continues, l'anova compare les moyennes entre groupes définis par le genre afin d'identifier les différences significatives. Les proxys identifiés lors de ces étapes constituent une base pour les analyses ultérieures.

Diagnostic des biais : Méthodes avancées

Deux outils complémentaires permettent d'évaluer l'impact des proxys sur les prédictions des modèles :

— **Valeur de Shapley** : Inspirée de la théorie des jeux coopératifs, cette méthode attribue une contribution marginale à chaque variable pour expliquer la prédiction globale. La formule générale est donnée par :

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Big[v(S \cup \{i\}) - v(S) \Big],$$

où S est un sous-ensemble de variables, $v(S)$ la performance du modèle utilisant S, et $|N|$ le nombre total de variables. Cette approche permet de quantifier précisément l'influence de chaque variable sur les prédictions et de repérer la place des proxys potentiels identifiés précédemment.

— **Importance par permutation** : Cette méthode évalue l'impact d'une variable sur les métriques d'équité en permutant ses valeurs de manière aléatoire. La dégradation des métriques d'équité indique le rôle joué par la variable dans la persistance des biais.

Méthodes de Mitigation : Orthogonalisation et Transport Optimal

Pour atténuer les biais dans les modèles prédictifs, deux techniques clés ont été mises en œuvre dans cette étude : l'orthogonalisation des variables explicatives et le transport optimal des distributions. Chaque méthode aborde les enjeux d'équité en réduisant la dépendance aux variables sensibles, soit lors de l'ingénierie des caractéristiques, soit en post-traitement.

Orthogonalisation des Variables Explicatives

L'orthogonalisation vise à réduire la corrélation entre les variables sensibles et les prédicteurs légitimes en décomposant chaque variable X_j en deux composantes : une partie indépendante de la variable sensible et une autre conservant sa corrélation. Cette transformation est définie par :

$$X_j^* = (1 - \alpha) X_j + \alpha X_j^{\perp}, \quad \alpha \in [0, 1],$$

où $X_j^{\perp}$ représente la composante de X_j orthogonale à la variable sensible. Le paramètre α contrôle le degré d'orthogonalisation, équilibrant équité et performance prédictive. Des valeurs élevées de α généralement améliorent l'équité mais peuvent dégrader la structure des données et la précision du modèle. Ce compromis souligne la nécessité de choisir α de manière adaptée au contexte d'application.

Transport Optimal des Distributions

La technique de transport optimal aborde l'équité au niveau des sorties des modèles en ajustant les prédictions conditionnelles Y pour les aligner avec une distribution cible, appelée «

barycentre » des groupes. Plus précisément, pour chaque groupe défini par une caractéristique sensible D, les distributions des prédictions sont modifiées afin de converger vers une distribution barycentrique indépendante de D. Cette approche de post-traitement garantit une parité démographique en minimisant la distance entre les distributions spécifiques aux groupes et le barycentre.

Ces méthodes de mitigation ont démontré leur efficacité dans les résultats expérimentaux, l'orthogonalisation réduisant les biais des proxys dans les variables explicatives et le transport optimal atteignant une parité démographique quasi parfaite. Cependant, chacune de ces approches implique des compromis qui doivent être soigneusement gérés pour maintenir transparence, performance prédictive et objectifs d'équité.

Résultats Principaux

Définitions et Mesures de l'Équité

Une analyse approfondie des métriques d'équité a permis de distinguer deux grandes visions de l'équité :

Équité de groupe : Cette approche vise à garantir une équité statistique entre des sous-groupes définis par des caractéristiques sensibles, telles que le genre ou l'origine ethnique. Elle repose sur des principes fondamentaux comme l'indépendance, la séparation ou la suffisance. Ces principes reposent sur l'indépendance statistique (conditionnelle ou non) et sont en général incompatibles.

Équité individuelle : Contrairement à l'équité de groupe, cette notion se concentre sur le traitement juste de chaque individu, indépendamment de son appartenance à un groupe spécifique. L'une des définitions les plus rigoureuses dans ce cadre est l'équité *contrefactuelle* ([5]). Elle stipule que les prédictions pour un individu devraient être identiques, que cet individu appartienne ou non à un autre groupe défini par la variable sensible. Cette équité repose sur la construction d'un **modèle causal** qui capte les relations entre les variables explicatives, la cible, et la caractéristique sensible. Un tel modèle permet de simuler des scénarios de « ce qui se serait passé si... », mais il est souvent complexe à établir en raison des hypothèses fortes qu'il nécessite, notamment sur l'identification des dépendances causales et sur la collecte de données fiables.

L'étude des définitions montre que la collecte des données sur les variables sensibles est essentielle pour mesurer et améliorer l'équité, ce qui soulève également des préoccupations en matière de confidentialité et de protection des données. Cela souligne la nécessité de concilier l'éthique et la confidentialité des données dans l'usage des systèmes d'IA.

Données utilisées

La base de données utilisée dans cette étude provient d'un portefeuille d'assurance automobile « responsabilité civile » d'un assureur belge en 1997. Elle contient les informations de **163 231 souscripteurs uniques**, chacun étant observé sur une période allant d'un jour à une année.

Les données sont organisées en plusieurs catégories :

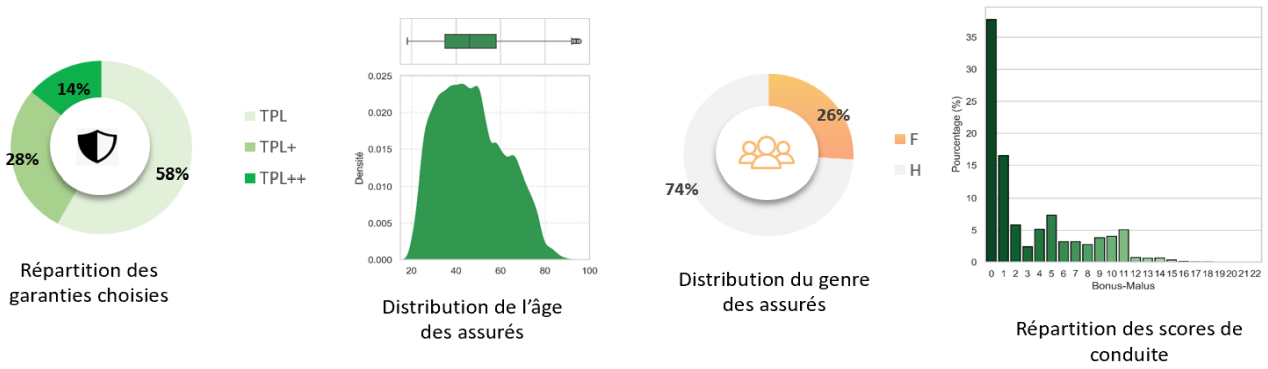

FIGURE 1 – Quelques caractéristiques dans les données

— **Variables d'intérêt** : Nombre de sinistres déclarés (`nclaims`), montant total des sinistres (`amount`) et durée d'exposition au risque (`exp`).

— **Caractéristiques des assurés** : Âge (`ageph`) et genre (`sex`).

— **Caractéristiques des contrats** : Type de couverture (`coverage`), niveau de bonus-malus (`bm`) et appartenance à une flotte (`fleet`).

— **Caractéristiques des véhicules** : Type de carburant (`fuel`), puissance (`power`), âge (`agec`) et usage (`use`, privé ou professionnel).

— **Données spatiales** : Coordonnées géographiques (`long`, `lat`) et code postal de la municipalité de résidence (`postcode`).

Cette base de données, issue du package R de **C. Dutang et A. Charpentier (2024) [12]**, a été utilisée dans plusieurs études. Elle ne présente pas d'anomalies notables et ne nécessite pas de traitements préliminaires majeurs. Une analyse descriptive révèle que plus de la moitié des contrats (58%) concernent exclusivement la garantie responsabilité civile simple (`TPL`), tandis que les couvertures complètes (`TPL++`) restent minoritaires (15).

Par ailleurs, les assurés sont majoritairement des hommes (78%), avec un âge principalement compris entre 25 et 75 ans et une moyenne située dans la quarantaine. Leur historique de conduite montre qu'un tiers des souscripteurs n'ont aucun accident enregistré dans le passé, comme en témoigne un score de bonus-malus égal à zéro (`bm` $= 0$).

Méthodes de Mitigation

Deux principales techniques de réduction des biais ont été évaluées :

— **Orthogonalisation des variables explicatives :**

Les résultats montrent une amélioration significative des métriques d'équité, comme la réduction de la distance de Kolmogorov de 45% pour les modèles de fréquence. Cependant, une orthogonalisation excessive peut dégrader la structure des données, entraînant une perte d'équité après un certain niveau d'orthogonalisation (*voir droite Fig. 4*).

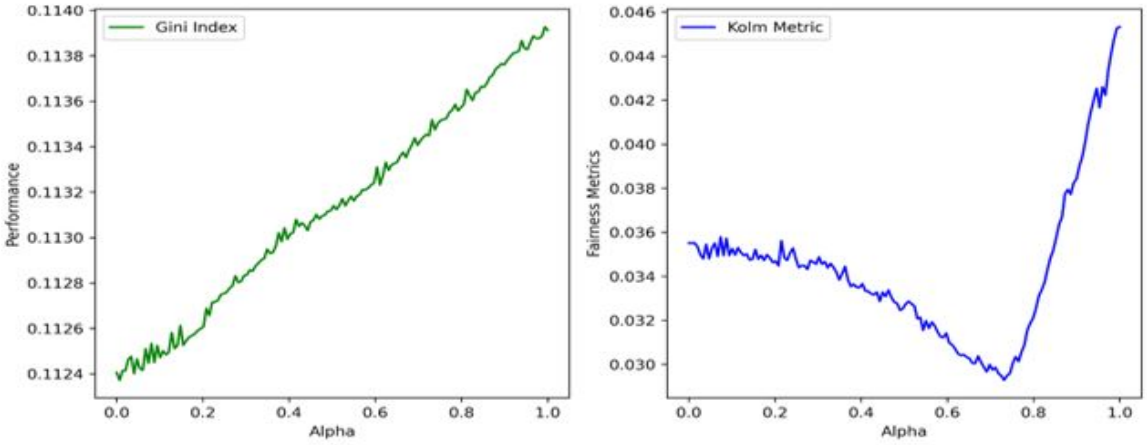

FIGURE 2 – Evolution du pouvoir prédictif et de l'équité du modèle de coûts selon le niveau d'orthogonalisation α.

— **Transport optimal des distributions :**

Les tests réalisés montrent une réduction de 80% de la distance de Kolmogorov par rapport à l'orthogonalisation, et une superposition des distributions de primes entre femmes et hommes (*voir Fig. 3*). Bien que très efficace, cette approche modifie légèrement les primes initiales, soulevant des questions sur l'acceptation par les assurés.

Analyse des Résultats

L'analyse révèle que la suppression des variables sensibles n'élimine pas totalement les biais, en raison de la présence de proxys. Par exemple, dans les modèles de coûts, des variables comme l'âge de l'assuré (`ageph`) et le niveau de score dans l'historique de conduite (`bm`) agissent comme proxys du genre, influençant indirectement les prédictions. D'où la nécessité d'utiliser des méthodes de réduction de discriminations avancées.

En outre, il est difficile de satisfaire simultanément plusieurs critères d'équité (de groupe). Les compromis entre équité et performance sont perceptibles surtout dans le cas où historiquement, la variable est un facteur de la cible. C'est le cas ici du modèle de fréquences où l'amélioration de l'équité s'accompagne d'une légère dégradation du pouvoir prédictif (mesuré par l'indice de Gini et de la déviance expliquée). C'est pourquoi, il est nécessaire de choisir judicieusement le principe d'équité à considérer en fonction de la situation rencontrée. La compréhension des définitions de l'équité est donc indispensable.

Perspectives et conclusion

Ce mémoire ouvre des pistes concrètes pour améliorer l'équité dans les modèles utilisés en assurance IARD. Cependant, plusieurs questions restent ouvertes :

— **Interprétabilité des modèles** : Comment garantir que les ajustements pour l'équité restent compréhensibles et transparents pour les utilisateurs finaux ? Les barycentres de Wasserstein par exemple semblent satisfaire par construction à cette exigence.

— **Extension à d'autres critères sensibles** : Bien que le genre ait été l'objet principal de cette étude, les mêmes analyses pourraient être appliquées à d'autres caractéristiques protégées, comme l'origine ethnique ou le niveau socio-économique. Comment formaliser le concept d'équité avec simultanément plusieurs variables sensibles est une question qui peut être examinée.

— **Impact sur le marché** : L'application des méthodes de mitigation, bien qu'efficace d'un point de vue éthique, pourrait affecter les dynamiques de souscription et la compétitivité des assureurs. Une analyse approfondie des implications commerciales est essentielle.

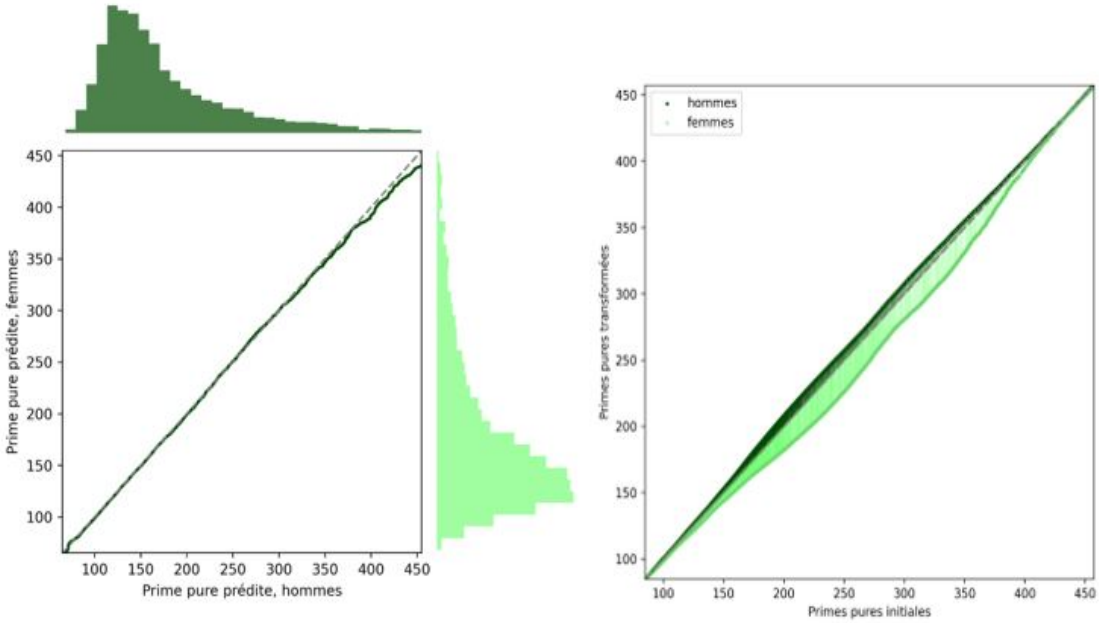

FIGURE 3 – Distributions conditionnelles de primes pures après mitigation (gauche), et comparaison entre prime pure initiale et prime pure équitable chez hommes et femmes (droite)

Ce mémoire met en évidence les défis et les opportunités liés à l'équité dans les modèles prédictifs d'assurance. Les méthodes proposées offrent des solutions concrètes pour réduire les biais tout en maintenant des performances acceptables. Cependant, leur mise en œuvre nécessite une attention particulière pour garantir un équilibre entre équité, efficacité et acceptation par les assurés. Ce travail constitue une contribution à l'évolution vers des systèmes d'assurance plus justes et éthiques et s'inscrit dans la continuité des mémoires précédents sur le sujet ([24], [30]).

Executive Summary

Context and Problem Statement

Artificial Intelligence (AI) is playing an increasingly significant role in the insurance sector, particularly in pricing and claims management. While promising in terms of efficiency and personalization, the use of predictive models raises significant concerns about fairness. These models can perpetuate or amplify biases present in the data, leading to indirect discrimination. This thesis addresses these challenges, with a particular focus on automobile insurance, and aims to propose methods to detect and mitigate such biases.

This study aligns with a strict European regulatory framework, including initiatives such as the **Directive 2000/78/EC** and the AI Act. These regulations impose obligations of non-discrimination, particularly regarding gender, and encourage the development of transparent and fair models.

Objectives and Methodology

The thesis focuses on three main objectives :

— **Understanding fairness principles** : Analyze the concepts of fairness through the literature in order to guide model designers and users towards definitions suited to their specific situations.

— **Measuring fairness in predictive models** : Evaluating the impact of biased data and proxies on algorithmic decisions using quantitative metrics such as statistical parity and the Kolmogorov distance.

— **Proposing mitigation methods** : Exploring solutions to mitigate biases, such as orthogonalization of explanatory variables and optimal transport of distributions.

The conceptual part is based on a collection of articles and books dedicated to defining fairness. It leads to the recommendation of a decision tree to guide the selection of a fairness definition based on the use case ([28]). The experimental part uses data from third-party liability automobile insurance in Belgium, provided by A. Charpentier and C. Dutang ([12]) in the R CASdatasets package.

Bias Diagnosis : Hypothesis Testing for Proxies

To identify potential proxies for the sensitive variable, statistical hypothesis tests were performed :

— **Correlation tests** : Correlations between explanatory variables and the sensitive variable (`DRIVER_GENDER`) were evaluated. These analyses help detect linear relationships that may introduce biases.

— **Independence tests** : Tests such as the **Kolmogorov-Smirnov test** or the **Chi-squared test** are used to verify whether distributions of explanatory variables differ significantly by gender.

— **Analysis of Variance (ANOVA)** : For continuous variables, ANOVA compares means between groups defined by gender to identify significant differences. Proxies identified during these steps form the basis for subsequent analyses.

Bias Diagnosis : Advanced Methods

Two complementary tools were used to evaluate the impact of proxies on model predictions :

— **Shapley Value** : Inspired by cooperative game theory, this method assigns a marginal contribution to each variable to explain the overall prediction. The general formula is :

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Big[v(S \cup \{i\}) - v(S) \Big],$$

where S is a subset of variables, $v(S)$ the model's performance using S, and $|N|$ the total number of variables. This approach quantifies each variable's influence on predictions, highlighting the role of potential proxies identified earlier.

— **Permutation Importance** : This method assesses the impact of a variable on fairness metrics by randomly permuting its values. The degradation in fairness metrics indicates the role played by the variable in maintaining biases.

Mitigation Methods : Orthogonalization and Optimal Transport

To address biases in predictive models, two key techniques were implemented in this study : the orthogonalization of explanatory variables and the optimal transport of distributions. Each method aims to enhance fairness by reducing dependence on sensitive variables, either during feature engineering or through post-processing.

Orthogonalization of Explanatory Variables

Orthogonalization reduces the correlation between sensitive variables and legitimate predictors by decomposing each variable X_j into two components : one independent of the sensitive variable and another retaining its correlation. This transformation is expressed as :

$$X_j^* = (1 - \alpha) X_j + \alpha X_j^{\perp}, \quad \alpha \in [0, 1],$$

where $X_j^{\perp}$ represents the component of X_j orthogonal to the sensitive variable. The parameter α controls the degree of orthogonalization, balancing fairness and predictive performance. Higher values of α improve fairness but may degrade the structure of the data and the model's accuracy. This trade-off underscores the need to select α appropriately based on the application context.

Optimal Transport of Distributions

The optimal transport technique addresses fairness at the level of model outputs by adjusting conditional predictions Y to align with a target distribution, referred to as the "barycenter" of the groups. Specifically, for each group defined by a sensitive attribute D, the prediction distributions are adjusted to converge toward a barycentric distribution independent of D. This post-processing approach ensures demographic parity by minimizing the distance between group-specific distributions and the barycenter.

These mitigation methods demonstrated their effectiveness in experimental results, with orthogonalization reducing proxy biases in explanatory variables and optimal transport achieving near-perfect demographic parity. However, each method involves trade-offs (especially the orthogonalization) that must be carefully managed to maintain transparency, predictive performance, and fairness objectives.

Key Results

Definitions and Fairness Metrics

An in-depth analysis of fairness metrics identified two primary perspectives on fairness :

Group Fairness : This approach aims to ensure statistical fairness between subgroups defined by sensitive characteristics such as gender or ethnicity. It relies on fundamental principles such as independence, separation, and sufficiency. These principles are based on statistical independence (conditional or unconditional) and are generally mutually incompatible.

Individual Fairness : Unlike group fairness, this concept focuses on the fair treatment of each individual, regardless of their group affiliation. One of the most rigorous definitions in this context is *counterfactual fairness* ([5]). It states that predictions for an individual should be identical, whether or not that individual belongs to another group defined by the sensitive variable. Counterfactual fairness relies on building a **causal model** that captures relationships between explanatory variables, the target, and the sensitive characteristic. Such a model enables simulation of "what-if" scenarios but is often complex due to strong assumptions about causal dependencies and data reliability. The study highlights that collecting data on sensitive variables

is essential for measuring and improving fairness. However, this raises concerns about privacy and data protection, emphasizing the need to balance ethics and confidentiality in AI systems.

Data Used

The database used in this study comes from a third-party liability automobile insurance portfolio of a Belgian insurer in 1997. It contains information on **163,231 unique policyholders**, each observed over a period ranging from one day to one year.

The data is organized into several categories :

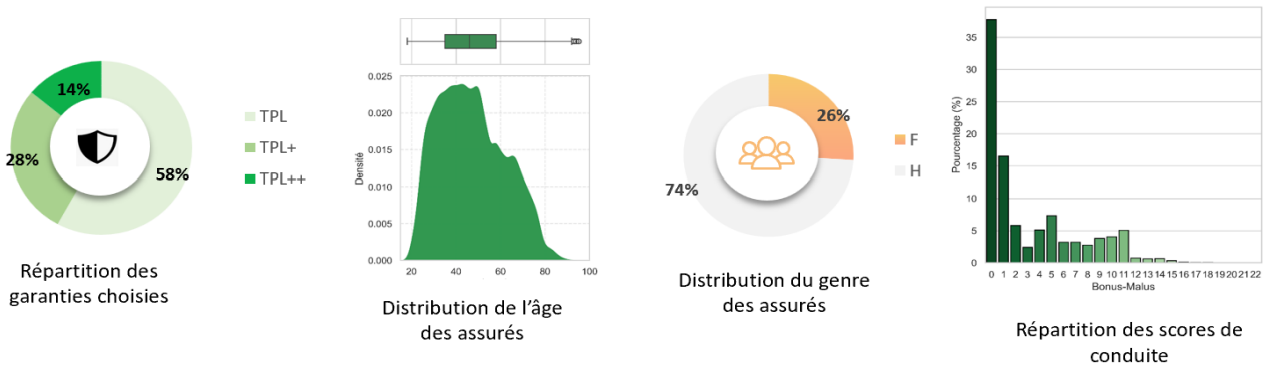

FIGURE 4 – Overview of some characteristics in the data

— **Key variables** : Number of claims filed (`nclaims`), total claim amount (`amount`), and exposure duration (`exp`).

— **Policyholder characteristics** : Age (`ageph`) and gender (`sex`).

— **Contract characteristics** : Coverage type (`coverage`), bonus-malus level (`bm`), and fleet membership (`fleet`).

— **Vehicle characteristics** : Fuel type (`fuel`), power (`power`), age (`agec`), and primary usage (`use`, private or professional).

— **Spatial data** : Geographic coordinates (`long`, `lat`) and postal code of the municipality of residence (`postcode`).

This database, sourced from the R package by **C. Dutang and A. Charpentier (2024) [12]**, has been used in multiple studies. It does not present any notable anomalies and does not require significant preprocessing.

A descriptive analysis reveals that more than half of the policies (58) cover only basic third-party liability (`TPL`), while comprehensive coverage (`TPL++`) remains less common (15).

Additionally, policyholders are predominantly male (78), with ages mostly ranging between 25 and 75 years, and an average age in the forties. Their driving history shows that one-third of policyholders have no recorded past accidents, as indicated by a bonus-malus score of zero (`bm` = 0).

Mitigation Methods

Two main bias mitigation techniques were evaluated :

— **Orthogonalization of explanatory variables :**

This method reduces the correlation between legitimate and sensitive variables. Results show significant improvements in fairness metrics, such as a 45% reduction in the Kolmogorov distance for frequency models. However, excessive orthogonalization can distort data structure, leading to reduced fairness after a certain threshold (*see Fig. 5*).

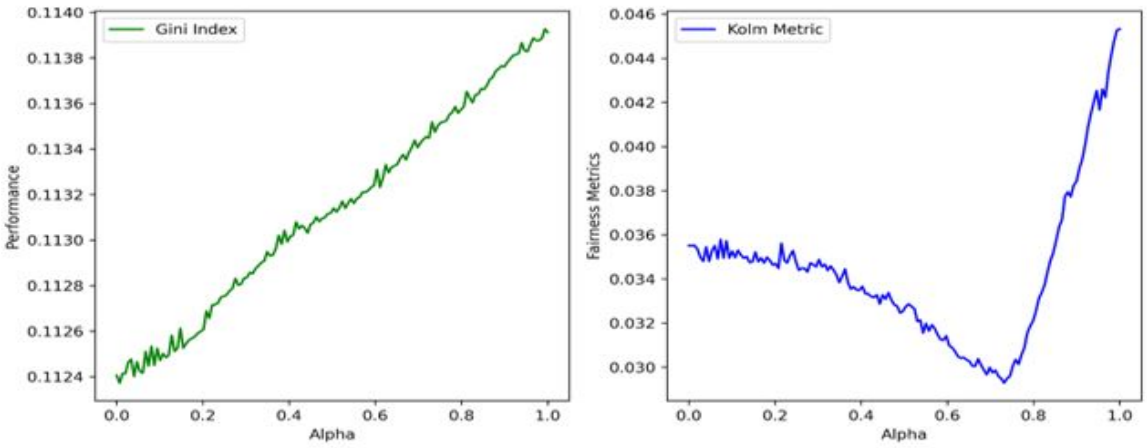

FIGURE 5 – Evolution of predictive power and fairness of cost models based on the level of orthogonalization α.

— **Optimal Transport of Distributions :**

This post-processing method adjusts predictions to ensure near-perfect demographic parity. Tests reveal an 80% reduction in the Kolmogorov distance compared to orthogonalization, with overlapping premium distributions between men and women (*see Fig. 6*). While highly effective, this approach slightly alters initial premiums, raising questions about policyholder acceptance.

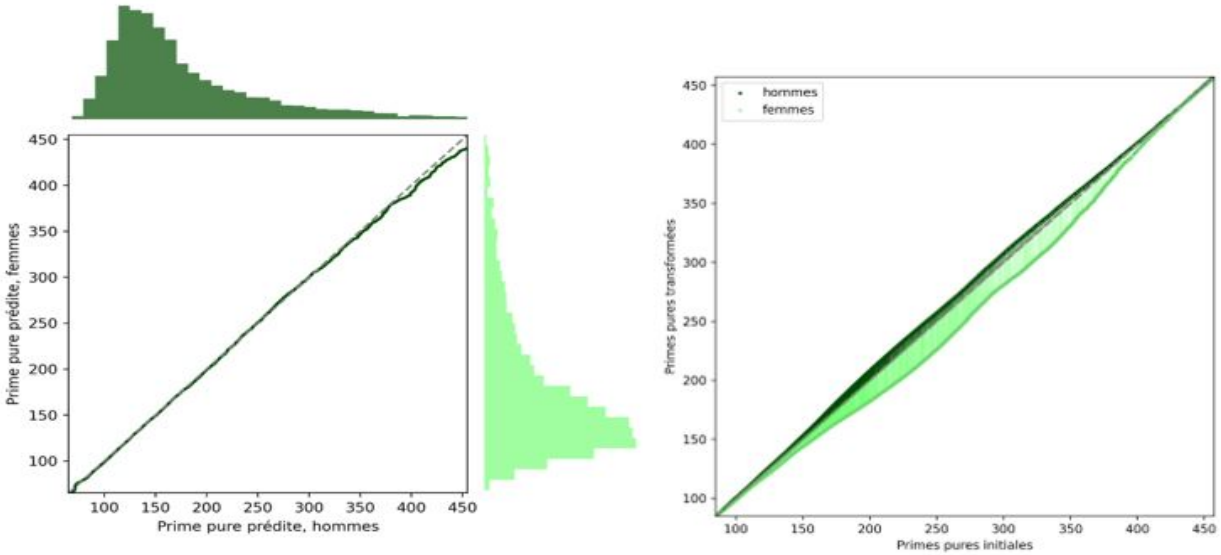

FIGURE 6 – Conditional distributions of pure premiums after mitigation (left) and comparison between initial pure premiums and equitable premiums for men and women (right).

Perspectives and Conclusion

This thesis provides concrete avenues to improve fairness in models used in non-life insurance. However, several questions remain :

— **Model interpretability** : How can fairness adjustments remain understandable and transparent for end-users ? Wasserstein barycenters, for example, seem to inherently satisfy this requirement.

— **Extension to other sensitive criteria** : Although gender was the primary focus of this study, similar analyses could be applied to other protected characteristics, such as ethnicity or socio-economic status. How to formalize fairness with multiple sensitive variables simultaneously is an area for further examination.

— **Market impact** : While effective from an ethical standpoint, applying mitigation methods could affect subscription dynamics and insurers' competitiveness. A deeper analysis of commercial implications is essential.

This thesis highlights the challenges and opportunities related to fairness in predictive models for insurance. The proposed methods provide concrete solutions to reduce biases while maintaining acceptable performance levels. However, their implementation requires careful attention to balance fairness, efficiency, and acceptance by policyholders. This work contributes to the evolution toward fairer and more ethical insurance systems, building on previous research ([24], [30]).

Table des matières

Introduction

L'intelligence artificielle (IA) occupe une place de plus en plus importante dans le secteur de l'assurance, où elle peut être utilisée pour automatiser des processus, évaluer les risques, et personnaliser les offres. Cette transformation technologique est porteuse de promesses, notamment en termes de gains d'efficacité et de personnalisation des offres. Cependant, elle s'accompagne également de défis éthiques majeurs, en particulier en matière d'équité. Les modèles prédictifs utilisés dans l'IA en général et le machine mearning en particulier peuvent perpétuer ou amplifier les biais présents dans les données, créant ainsi des discriminations parfois subtiles mais significatives.

L'évolution des réglementations européennes, comme la directive sur la tarification non genrée ou encore le projet d'IA Act, souligne l'importance de l'équité dans le cadre des modèles algorithmiques. Ces cadres législatifs visent à prévenir les discriminations directes et indirectes tout en encourageant le développement de solutions transparentes et éthiques.

Ce mémoire s'inscrit dans cet enjeu sociétal et scientifique, avec pour objectif principal d'analyser, de mesurer et de mitiger les biais dans les modèles prédictifs appliqués au domaine assurantiel, et plus particulièrement à l'assurance automobile. Les travaux se concentrent sur trois axes majeurs :

1. **Définitions et principes d'équité** : Une revue approfondie de la littérature est proposée afin de clarifier les concepts d'équité et de discrimination dans un cadre actuariel et algorithmique ;
2. **Évaluation des biais** : À l'aide de données réelles et simulées, les biais des modèles sont mesurés à travers des métriques comme la différence de parité statistique ou la distance de Kolmogorov. L'importance des proxys dans la persistance des biais est également analysée ;
3. **Proposition de solutions** : Des méthodes telles-que l'orthogonalisation des variables explicatives ou le transport optimal des distributions, sont explorées pour réduire les biais tout en maintenant des performances acceptables.

Le mémoire est structuré en cinq chapitres. Après une introduction, le premier chapitre définit l'équité en assurance et ses enjeux. Le deuxième et le troisième chapitre abordent les mesures de l'équité et les méthodes de mitigation des biais, tandis que les deux derniers appliquent ces concepts à l'assurance automobile, en analysant les biais des modèles et les solutions pour les corriger.

Les résultats obtenus révèlent des compromis nécessaires entre équité, performance et interprétabilité des modèles. Ce mémoire ouvre des perspectives pour le développement de systèmes d'assurance plus justes et éthiques, tout en posant des questions cruciales sur leur acceptabilité commerciale et sociétale.

Chapitre 1

L'équité en assurance : intérêt et enjeux

L'équité est un concept universellement reconnu, souvent associé à des valeurs fondamentales telles que la justice, l'impartialité et l'égalité. Selon le dictionnaire Oxford, elle se définit comme « un traitement ou un comportement parfaitement juste, sans favoritisme ni discrimination ». Ce principe joue un rôle crucial dans divers contextes, y compris dans le domaine de l'assurance. Au-delà de son application légale et sociétale, l'équité s'impose de plus en plus dans les secteurs à forte intensité de données, où elle est essentielle pour prévenir les biais et les discriminations. Ce chapitre explore le concept d'équité, ses implications en assurance, ainsi que les défis liés à son application dans un secteur de plus en plus influencé par les technologies numériques et l'intelligence artificielle.

1.1 Comprendre l'équité : définitions et concepts clés

Le terme "équité" est emprunté du latin classique "aequitas" qui peut signifier selon les situations égalité, juste proportion ou encore esprit de justice. Dans son usage courant, l'équité est assimilée à la capacité à accorder à chacun ce qui lui est dû (Dictionnaire de l'académie française). Derrière cette définition linguistique simple se cache une multitude de points de vue et de considérations de l'équité.

1.1.1 L'équité et la discrimination : définitions fondamentales

L'équité, bien qu'universellement valorisée, est un concept multidimensionnel. En plus de sa définition selon le dictionnaire Oxford, elle est souvent décrite comme la capacité à garantir un équilibre entre les droits et les responsabilités des individus, ou encore à éviter des traitements inéquitables. Ce principe est intrinsèquement lié à la notion de discrimination, définie comme une action ou une pratique qui exclut, désavantage ou différencie simplement des individus ou des groupes d'individus sur la base d'une caractéristique attribuée ou perçue »[16]. Le droit européen, le principe de non-discrimination interdit les différences de traitements basés sur un ensemble de critères dits protégés.

Par ***caractéristique protégée*** on parle d'une caractéristique identifiable, objective ou personnelle, ou une situation, par laquelle des individus ou des groupes se distinguent les uns des autres. C'est une propriété jugée non pertinenete pour justifier des différences de traitement ou l'octroi d'un avantage particulier. Le manuel de droit européen en matière de non discrimination [22] cite comme caractéristiques protégées le Sexe, la race, la couleur, la langue, la religion, les opinions politiques ou toutes autres opinions, l'origine raciale ou sociale, l'appartenance à

une minorité nationale ou encore la fortune.

Les discriminations se classent généralement en deux grandes catégories : **discrimination directe** et **discrimination indirecte**. Ces distinctions permettent de mieux comprendre les mécanismes d'injustice dans différents contextes.
Une **discrimination directe** se produit lorsqu'une personne ou un groupe est traité de manière moins favorable qu'une autre personne ou un autre groupe dans une situation comparable, directement en raison d'une caractéristique protégée. Par exemple, refuser l'accès à une assurance automobile à une femme simplement en raison de son sexe constitue une discrimination directe. À l'inverse, une **discrimination indirecte** survient lorsqu'une règle ou une pratique apparemment neutre entraîne, en réalité, un désavantage disproportionné pour des individus possédant une caractéristique protégée, à moins que cette règle ne puisse être justifiée objectivement par un objectif légitime et que les moyens pour l'atteindre soient appropriés et nécessaires. Par exemple, exiger des preuves de résidence de longue durée pour bénéficier de certaines prestations d'assurance pourrait indirectement exclure des populations immigrées, même si cette exigence n'évoque aucune appartenance ethnique ou origine.

Ces distinctions soulignent la complexité de la notion de discrimination et la nécessité d'analyser attentivement les pratiques et politiques pour s'assurer qu'elles ne génèrent pas d'effets inéquitables, intentionnels ou non.

1.1.2 L'équité d'un point de vue légal, sociologique et philosophique

L'équité, en tant que concept multidimensionnel, se révèle complexe à appréhender et à définir. Sur le plan légal, l'équité se traduit par la mise en œuvre de principes de justice et de non-discrimination. Cependant, comme le souligne Vandenhole (cité par A. Charpentier [5], p4) il n'existe pas de définition universellement acceptée de la discrimination, ce qui complique l'application de ces principes dans la pratique juridique. En Europe, les directives antidiscriminatoires telles que la Directive 2000/78/CE, établissent un cadre général en faveur de l'égalité de traitement en matière d'emploi, de prévoyance et sécurité sociale, d'éducation, d'accès à la justice et à la fourniture des biens et services. Cependant, malgré leur ambition louable, ces directives se heurtent fréquemment à des difficultés d'application et à des interprétations divergentes au sein des différents États membres de l'Union européenne .
D'un point de vue philosophique, l'équité est souvent associée à la notion d'égalité des chances. John Rawls ([27]), dans sa "Théorie de la justice" (, propose que chaque individu devrait avoir une chance équitable de réussir, indépendamment de ses conditions de départ. Cette approche met en avant l'idée que les talents et la motivation ne sont pas répartis de manière égale parmi les différentes populations, ce qui nécessite des mesures correctives pour assurer une véritable égalité des opportunités .
Sur le plan sociologique, l'équité est souvent étudiée à travers les conséquences de la discrimination systémique et des inégalités structurelles. Les normes et règles sociales peuvent produire des résultats disproportionnellement défavorables pour certains groupes, même en l'absence d'intention discriminatoire. Ainsi, les sociologues s'intéressent à la manière dont les politiques et les pratiques institutionnelles peuvent perpétuer des inégalités, et cherchent des solutions pour promouvoir une plus grande justice sociale.
Sous ses différentes facettes sociales, légales et philosophiques, l'équité peut jouer un rôle essentiel dans la quête de justice au sein de nos sociétés. Cependant, lorsqu'il s'agit de l'application de ce principe dans des domaines spécifiques, tels que l'assurance, l'équité revêt un sens encore plus précis et technique. Lorsqu'il est évoqué en assurance, sa signification première est particulière, car il se trouve au cœur même de ce secteur. Voyons maintenant à quoi renvoie ce principe dans la pratique actuarielle.

FIGURE 1.1 – Nuage de mots proches de l'équité sur des sites institutionnels [1].

1.1.3 L'équité actuarielle

En assurance, chaque assuré est protégé contre des risques définis en échange du paiement d'une prime. Le principe d'équité actuarielle stipule que la prime (pure) payée par chaque client doit réflèter le risque qu'il apporte au portefeuille. Autrement dit, chaque client est chargé conformément au de niveau de ses pertes financières potentielles. A titre d'illustration, imaginons une situation où deux personnes souhaitent souscrire une assurance automobile. La première personne, qui a 25 ans et conduit en ville, a un risque annuel d'accident estimé à 10 %. La seconde personne, âgée de 50 ans et conduisant principalement en zone rurale, a un risque annuel d'accident de 5 %. Le principe de l'équité actuarielle voudrait que les primes soient proportionnelles à ces risques. Si l'assureur estime que le coût moyen d'un accident est de 1 000 euros, la première personne paiera une prime annuelle de 100 euros (10 % de 1 000 euros), tandis que la seconde paiera 50 euros (5 % de 1 000 euros). Ainsi, chacun paie en fonction de son propre risque.

L'équité actuarielle est mise en application lors de la souscription, processus consitant à examiner, à accepter ou à rejeter les risques d'assurance et à classer ceux qui ont été sélectionnés, afin de facturer la prime appropriée pour chacun d'entre eux (Harvey Rubin, cité par X. Landes [19]). Ceci évite que certains assurés supportent les risques des autres de manière injustifiée. De plus, cette approche permet de minimiser les phénomènes indésirables tels-que l'aléa moral ou la sélection adverse par lesquels les pertes potentielles sont beaucoup plus importantes que les primes collectées. L'équité actuarielle contribue ainsi à la stabilité financière du système d'assurance.

Par l'équité actuarielle, les individus doivent payer pour leurs propres risques et uniquement pour leurs propres risques, ce qui imprime à ce principe des marques de justice. Cependant, est-ce vraiment un principe juste ? Dans la troisième partie de son article "How fair is actuarial fairness" [19], Xavier Landes soulève une objection à l'équité actuarielle (entre autres) par rapport à la propriété du risque. Il indique qu'être porteur d'un risque n'implique pas forcément en

1. .gouv.fr ou uropa.eu OR ou academie-francaise.fr ou cnrtl.fr

être responsable. Tel est généralement le cas lorsqu'un facteur de risque correspond à une caractéristique acquise à la naissance (l'origine, le genre, le gênes). En assurance vie par exemple, le genre est reconnu comme facteur de mortalité. Struyck (1972), cité par A. Charpentier ([5], P.36) montre qu'à l'âge de 20 ans l'espérance de vie est d'environ 5 années plus faible chez les hommes que chez les femmes. Même si cette différence peut être attribuable au style de vie (comportements à risque, suivi médical, stress et charge mentale), est-elle le résultat d'un choix ? De façon plus générale, est-il juste que les primes d'assurance dépendent de facteurs pour lesquels les assurés ne sont pas directement responsables ? Ces interrogations permettent de se rendre compte du besoin d'équité en assurance, au-delà de l'équité actuarielle.

1.2 Intérêt de la recherche d'équité dans les tarifs d'assurance

La section précédente a posé la question de la justice des primes découlant du principe d'équité actuarielle, en soulignant les limites de ce principe lorsque des facteurs comme le genre, ou toute autre caractéristique non liée aux choix individuels de l'assuré, jouent un rôle déterminant. Dans le secteur de l'assurance, les exigences en matière d'équité vont au-delà de l'équité actuarielle traditionnelle. Les assureurs sont désormais tenus de proposer des tarifs qui ne dépendent pas directement de ces facteurs sensibles, une exigence façonnée par un cadre réglementaire strict et par les évolutions technologiques.

En effet, l'adoption croissante des outils d'intelligence artificielle (IA) transforme profondément l'industrie de l'assurance. Ces technologies, largement utilisées pour améliorer la gestion des sinistres, la détection des fraudes ou encore la tarification, apportent des gains significatifs en termes d'efficacité et de personnalisation. Cependant, elles introduisent également des risques notables, notamment l'apparition ou l'amplification de biais dans les modèles utilisés. Ces biais peuvent survenir à différents stades du cycle de vie des outils d'IA et soulèvent des questions éthiques et sociétales cruciales.

Par ailleurs, le cadre réglementaire européen impose des exigences strictes en matière de non-discrimination, notamment avec la *Directive 2000/78/CE*, qui interdit toute distinction basée sur le genre dans la tarification des produits assurantiels. Ces règles renforcent la nécessité de développer des modèles transparents, responsables et alignés sur des principes éthiques robustes.

Enfin, au-delà des considérations légales et technologiques, le risque réputationnel constitue une autre préoccupation majeure pour les assureurs. Une gestion perçue comme inéquitable des tarifs d'assurance peut entraîner une perte de confiance de la part des clients et des parties prenantes, avec des impacts significatifs sur la compétitivité des entreprises.

1.2.1 L'utilisation croissante des outils d'IA en assurance

L'intelligence artificielle joue un rôle de plus en plus important dans le secteur assurantiel, en automatisant et en optimisant des processus stratégiques. Deux exemples illustrent cette tendance :

— **Gestion des sinistres (acceptation automatique)** : Les algorithmes d'IA sont utilisés pour évaluer automatiquement les demandes d'indemnisation, réduisant ainsi les délais de traitement. Ces modèles, en analysant rapidement les informations fournies par les assurés, permettent une prise de décision quasi instantanée, augmentant l'efficacité opérationnelle.

— **Détection des fraudes** : Les outils basés sur le machine learning identifient des schémas anormaux dans les réclamations. Par exemple, ils peuvent repérer des incohérences dans les informations fournies ou des similitudes avec des cas de fraude connus, facilitant ainsi la prévention proactive des fraudes.

— **Tarification basée sur l'IA** : Les algorithmes de machine learning exploitent de vastes ensembles de données, incluant l'historique des sinistres, le profil des assurés et des variables externes (conditions météorologiques, tendances économiques). Cette approche peut améliorer la précision des modèles actuariels et permettre une tarification plus individualisée et réactive aux évolutions du risque.

Ces applications démontrent le potentiel transformateur de l'IA dans l'assurance. Cependant, leur efficacité et leur impartialité dépendent de la qualité des données, des modèles et des processus qui sous-tendent leur développement.

Le cycle de développement des outils d'IA et les types de biais

Le développement des systèmes d'intelligence artificielle (IA) dans des domaines comme l'assurance ou la finance implique plusieurs étapes : la collecte des données, la création des modèles, et leur utilisation. À chaque étape, des biais spécifiques peuvent se manifester, menaçant l'équité et l'efficacité des décisions prises. Ces biais résultent souvent d'une combinaison de limitations dans les données, de choix techniques dans les modèles, ou d'interactions humaines avec ces outils.

Un biais peut être défini comme une distorsion systématique dans la prise de décision ou la prédiction, qui résulte d'une représentation inadéquate des informations disponibles. Ces biais peuvent affecter l'équité entre différents groupes ou individus, compromettant ainsi la fiabilité des résultats.

1. Collecte et préparation des données

Les biais introduits à ce stade trouvent leur origine dans la nature des données utilisées pour entraîner les modèles :

— **Biais de mesure** : Ils surviennent lorsque les variables collectées ne sont pas correctement définies ou mesurées. Cela conduit à des représentations inexactes des phénomènes étudiés.
Exemple : En assurance auto, utiliser le code postal comme indicateur de risque peut introduire une discrimination géographique, car certaines zones pourraient refléter indirectement le revenu ou la composition ethnique de la population.

— **Biais de représentation** : Ce biais se produit lorsque les données ne reflètent pas correctement la population cible. Les décisions basées sur ces données peuvent alors désavantager certains groupes.
Exemple : Une base de données de sinistres majoritairement issue d'assurés urbains risque de sous-estimer les risques associés aux conducteurs en zones rurales.

— **Biais d'échantillonnage** : Ce biais apparaît lorsqu'un groupe de la population est surreprésenté ou sous-représenté dans les données.
Exemple : Si les jeunes conducteurs novices sont sous-représentés dans une base de données d'assurance auto, les modèles pourraient ne pas prédire correctement leur risque élevé de sinistres.

2. Création des modèles

Lors de la construction des modèles prédictifs, des choix techniques peuvent introduire des biais, même si les données utilisées sont fiables :

— **Biais algorithmique** : Ce biais provient des hypothèses faites lors de la conception des algorithmes, comme les fonctions d'optimisation ou les critères de performance.

Exemple : Une fonction de coût qui minimise globalement les erreurs peut ignorer les performances pour des groupes spécifiques, tels que les jeunes conducteurs.

— **Biais d'omission** : Il se manifeste lorsque des variables pertinentes sont exclues du modèle, soit par erreur, soit en raison de contraintes techniques ou éthiques.
Exemple : En assurance santé, ne pas inclure des indicateurs d'habitudes alimentaires peut sous-estimer les risques liés à certaines maladies.

— **Biais d'agrégation** : Il apparaît lorsqu'on analyse les données de manière globale sans tenir compte des différences entre sous-groupes.
Exemple : Regrouper les conducteurs avec et sans antécédents de sinistres peut fausser les prévisions de primes en masquant les risques spécifiques à chaque groupe.

3. Utilisation et interaction avec les modèles

Même après le développement, des biais peuvent émerger lors de l'utilisation des modèles dans un contexte réel :

— **Biais de popularité** : Ce biais survient lorsque certains résultats ou décisions deviennent dominants en raison de leur visibilité accrue.
Exemple : Dans un système de scoring, si certains produits financiers sont fréquemment proposés, cela pourrait limiter l'accès des clients à des options potentiellement plus adaptées.

— **Biais émergents** : Ces biais se manifestent lorsque le modèle interagit avec des populations ou des environnements différents de ceux pour lesquels il a été conçu.
Exemple : Un modèle d'évaluation du risque auto conçu pour un pays européen pourrait être moins pertinent dans un pays où les conditions routières et les comportements de conduite diffèrent.

— **Biais d'évaluation** : Ce biais est lié à l'utilisation de benchmarks inadéquats pour évaluer les performances des modèles.
Exemple : Évaluer un modèle de prédiction des sinistres en assurance habitation sur des données limitées à des zones urbaines peut biaiser les conclusions pour les zones rurales.

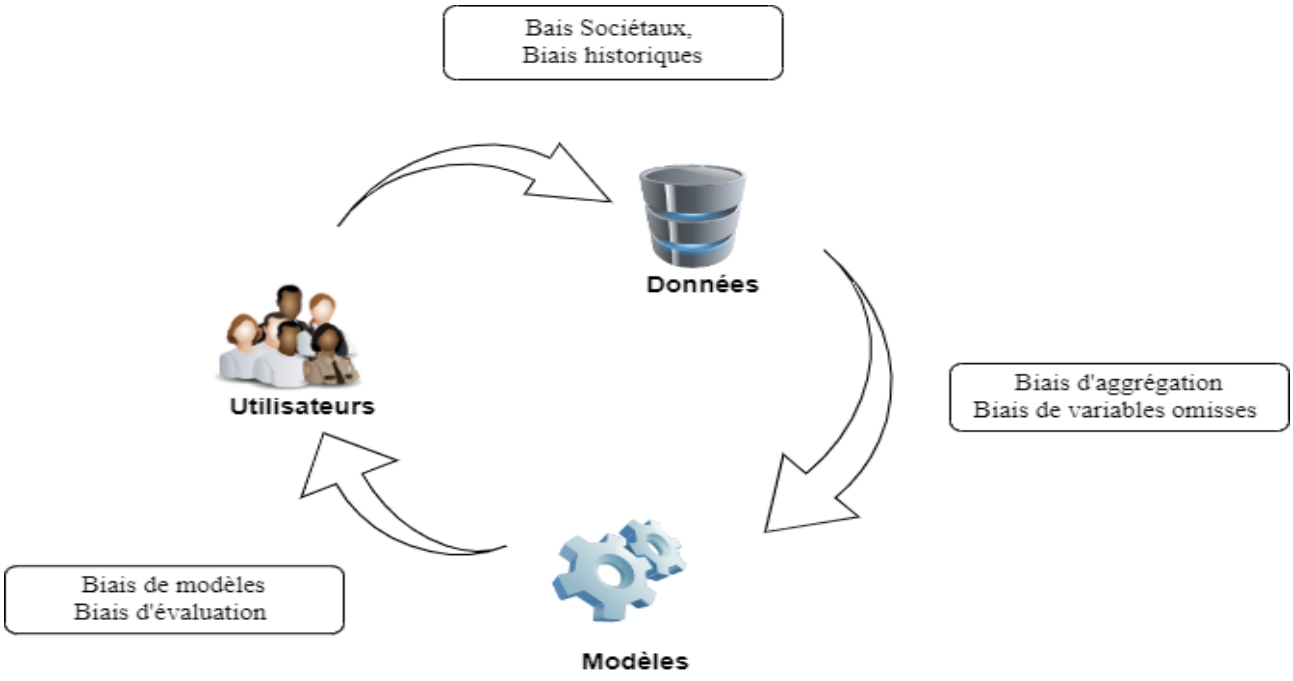

FIGURE 1.2 – Cycle des biais, sources potentielles de discriminations des modèles [2].

L'adoption des outils d'IA dans l'assurance offre des opportunités considérables pour améliorer l'efficacité et réduire les coûts. Cependant, ces outils comportent des risques importants de biais, qui peuvent affecter à la fois la qualité des résultats et la perception d'équité par les assurés. Une approche rigoureuse est donc essentielle pour minimiser ces biais et garantir que les technologies d'IA contribuent à une plus grande équité, tout en respectant les contraintes réglementaires et éthiques.

1.2.2 Le paysage réglementaire Européen du secteur assurantiel en matière d'équité.

Lorsque l'on aborde les questions d'équité et, plus largement, d'éthique dans le cadre de l'assurance au sein de l'Union européenne, deux réglementations majeures sont immédiatement évoquées.

— **La directive Européenne sur la tarification non genrée**

Il s'agit d'une réglémentation adoptée par la Commission Européenne rendue obligatoire en fin 2012 pour tout assureur sur le marché. L'institution reconnaît comme discriminatoires et donc incompatibles avec la charte des droits fondamentaux de l'UE les différences de primes d'assurance basées uniquement sur le genre. De ce fait, les compagnies d'assurance sont sommées d'appliquer des tarifs non genrées sur leurs produits d'assurance. Ceci signifie que des clients ayant les mêmes caractéristiques paient des prix identiques pour les mêmes produits indépendamment de leur genre ([7], [8]).

— **Les lignes directrices en matière d'éthique pour une IA digne de confiance** [9]

Elles ont été élaborées dans un contexte où l'intelligence artificielle (IA) prend une place de plus en plus importante au sein des sociétés européennes. Pour répondre aux enjeux que cela représente, un groupe d'experts indépendants, mandaté par la Commission européenne en juin 2018, a été chargé de formuler des recommandations précises. Leur objectif principal est de veiller à ce que l'IA soit développée de manière éthique, robuste et respectueuse des valeurs fondamentales de la société. Ce cadre est destiné à guider les entreprises et organisations, qu'elles soient en Europe ou qu'elles opèrent sur le marché européen, dans leur utilisation des technologies d'IA.

L'idée est de s'assurer que les systèmes d'IA respectent non seulement la législation en vigueur, mais aussi les principes éthiques, tout en étant techniquement fiables et socialement responsables. Les principes éthiques sur lesquels ces lignes directrices mettent l'accent incluent le respect de l'autonomie humaine, la prévention des risques, la transparence des décisions et l'équité. En ce qui concerne l'équité, il est souligné que les utilisations de l'IA doivent traiter de manière juste, sans discrimination ni stigmatisation les groupes et les individus de manière à répartir également les bénéfices et les coûts. Ainsi, les systèmes d'IA doivent éviter de reproduire des biais ou des discriminations existantes.

Néanmoins, il est important de noter que ces lignes directrices restent relativement générales quant à la manière de concrétiser l'équité, en particulier dans des secteurs spécifiques comme celui de l'assurance. Bien qu'elles offrent un cadre de référence solide et des principes directeurs, elles ne prescrivent pas de méthodes précises pour garantir cette équité dans des domaines spécifiques. Les entreprises peuvent donc adapter ces recommandations en fonction de leur contexte opérationnel.

2. Inspiré de Ninareh M. et al. (2022)[23]

1.2.3 Le risque réputationnel

Le risque réputationnel est un enjeu critique pour les compagnies d'assurance, influençant non seulement leur image publique mais aussi leur viabilité à long terme. Ce type de risque se définit comme la menace ou la perte potentielle d'une entreprise due à une atteinte à sa réputation, qui peut résulter de diverses actions perçues comme injustes, discriminatoires, ou non éthiques par le public ou les parties prenantes. Dans le contexte de l'assurance, ce risque peut émerger de pratiques perçues comme inéquitables, que ce soit dans la tarification des primes ou dans la gestion des sinistres.

Les origines du risque réputationnel en assurance sont multiples. Elles peuvent inclure la perception publique de discrimination, par exemple, si une compagnie est vue comme favorisant certains groupes au détriment d'autres sans justification claire et transparente. De plus, dans un environnement où l'intelligence artificielle (IA) et les algorithmes sont de plus en plus utilisés pour déterminer les primes et gérer les sinistres, la méfiance du public à l'égard de ces technologies peut exacerber ce risque. Si les décisions prises par des systèmes d'IA sont perçues comme opaques, injustes, ou biaisées, la réaction du public peut gravement nuire à la réputation de l'assureur.

Ce risque réputationnel devrait pousser les assureurs à rechercher davantage d'équité dans leurs pratiques, non seulement pour se conformer aux exigences réglementaires mais aussi pour maintenir la confiance des consommateurs. En ce qui concerne la tarification, cela se traduirait par une approche plus transparente et justifiable des critères utilisés pour déterminer les primes. Par exemple, les assureurs peuvent être incités à adopter des modèles tarifaires qui excluent les facteurs perçus comme discriminatoires, comme le genre ou l'origine ethnique (Gender Directive), en faveur de critères plus objectifs et vérifiables. De même, dans la gestion des sinistres, un effort supplémentaire peut être fait pour garantir que les décisions sont prises de manière équitable et cohérente, en évitant toute apparence de partialité ou d'injustice.

En fin de compte, l'atténuation du risque réputationnel par la recherche d'équité n'est pas seulement une question de conformité réglementaire, mais également une stratégie essentielle pour préserver la confiance du public et assurer la pérennité des activités de l'assureur. Les compagnies qui échoueront à intégrer ces considérations dans leurs opérations courent non seulement le risque de sanctions réglementaires, mais aussi de perdre la confiance de leurs clients, ce qui pourrait avoir des conséquences financières significatives.

Chapitre 2

Définitions et mesures de l'équité des modèles prédictifs

La mesure et la réduction des discriminations dans les modèles d'apprentissage ou d'IA en général nécessitent une définition claire de ce qu'est l'équité ou l'absence de dicriminations. Le précédent chapitre a fourni une vision conceptuelle de l'équité, sans en donner une version quantifiable ou mesurable. Il existe plusieurs définitions mathématiques de l'équité dans la littérature, et ce chapitre en présente les grandes lignes. Les métriques d'évaluation de l'équité des modèles y sont également définies et même évaluées. Avant d'y parvenir, il est nécessaire de poser le cadre dans lequel s'inscrivent ces notions et grandeurs.

2.1 Cadre et notations pour les définitions de l'équité

Dans la littérature, plusieurs définitions de l'équité sont formulées, selon les situations et les objectifs poursuivis. Le point commun de ces définitions est qu'elles requièrent la présence - dans la population cible du modèle évalué - de plusieurs catégories d'individus pour lesquelles l'équité est définie. Ces catégories sont représentées par une ou plusieurs variables catégorielles dites "sensibles" ou "protégées". Le genre, l'origine, la zone de résidence sont des exemples courants de variables sensibles rencontrées dans le domaine de l'assurance. La variable sensible (ou combinaison des variables sensibles) est notée D et considérée binaire par la suite, prenant ses valeurs dans l'ensemble $\mathcal{D} = \{d_1, d_2\}$. Cette considération peut se comprendre dans la mesure où toutes les situations peuvent se ramener à distinguer 2 catégories d'individus, celle des "favorisés" et celle des "défavorisés".

Le cadre est celui d'un modèle d'apprentissage supervisé, inspiré de celui décrit par Lindholm M. et al. (2024) [21]. Les individus sont décrits, en plus de la variable sensible, par d'autres variables représentées par un vecteur aléatoire connu X et leurs caractéristiques d'intérêts représentées par une variable aléatoire Y (la target), binaire ou continue. Les variables (ou features) X seront qualifiées « non-sensibles »ou « légitimes ». L'ensemble des valeurs possibles de X est noté $\mathcal{X}$ et celui de Y $\mathcal{Y}$.
Le triplet (D, X, Y) est porté par un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, où $\mathbb{P}$ est la mesure de probabilité monde-réel. Pour tenir compte de la possibilité d'utiliser ou pas la variable sensible, on note $Z = X \oplus (XUD)$ [1].

— **Cas d'une variable d'intérêt binaire**

1. Le symbole $\oplus$ renvoie au "OU" exclusif. Ici Z représente donc les features/variables non sensibles, ou bien la combinaison variable sensible et variables non-sensibles

Le modèle prédictif est un classificateur binaire. Il est noté $\hat{Y} = \mathbf{1}_{m(Z)\geq\tau} = m_\tau(Z)$ où $m(Z)$ est le score prédit, et τ un seuil. Comme exemples, on peut citer les arbres de décisions, les forêts aléatoires et les régressions logistiques.

— **Cas d'une variable d'intérêt continue**

Le modèle prédictif est un régresseur, noté simplement : $\hat{Y} = m(Z)$. Les exemples précédents avec seul le score considéré peuvent être cités ici.

Une fois le cadre posé, les définitions de l'équité peuvent être formulées aisément.

2.2 Définitions et mesures de l'équité des modèles : état de l'art

Dans la littérature, les définitions probabilistes de l'équité se regroupent en deux catégories : celles qui s'intéressent aux groupes, et celles qui se focalisent sur les individus. Chacune de ces approches traite la question de façon particulière et mérite qu'on s'y intéresse.

2.2.1 L'équité au niveau individuel

L'équité d'un modèle selon cette approche est définie au travers de critère vérifié sur chaque individu ciblé par ce modèle. Le premier critère proposé dans la littérature (A. Charpentier [5], Alessandro C. et al [4]) est que les individus similaires - au sens de leurs caractéristiques non sensibles uniquement - doivent être traités, dans des circonstances similaires, de la même manière par le modèle. Le second critère d'équité individuelle repose sur la connaissance des relations causales entre les variables en présence. L'exigence est alors que chaque individu ait la même prédiction que celle qu'il aurait obtenue s'il appartenait à une catégorie de D différente de l'actuelle.

Définitions basées sur un critère de similitude

Une manière simple de traiter de manière identiques les individus similaires au sens des variables non-sensibles est de retirer la variable sensible du modèle. De ce fait, les individus ayant les mêmes features auront les mêmes prédictions. On parle alors d' **équité par inconscience** ou **équité par ignorance** (en anglais *fairness through unawareness - FTU*)[5], [4]. C'est l'idée derrière la Directive Européenne sur la tarification non genrée par exemple (voir section 1.2.2). Cependant, cette approche est ***inconsciente*** du fait que certaines variables non-sensibles peuvent porter l'information sur D. Ainsi, même lorsqu'elle n'est pas utilisé explicitement dans un modèle, elle peut l'être implicitement au travers de certaines features X. C'est pourquoi, d'autres définitions sont proposées.

Une autre définition basée sur la similitude est la *propriété de Lipschitz*, proposée par Dwork et al. (2011). Un modèle ($\hat{Y} = m(Z)$) vérifie cette propriété si pour toute paire d'individus $\{i, j\}$, une distance faible dans leurs features induit une distance faible dans leurs prédictions. Formellement :

$$D_{\mathcal{Y}}(\hat{Y}_i, \hat{Y}_j) \leq \gamma D_{\mathcal{X}}(X_i, X_j), \gamma \geq 0, \forall i \neq j \text{ où } D_{\mathcal{X}} \text{ et } D_{\mathcal{Y}} \text{ distances sur } \mathcal{X} \text{ et } \mathcal{Y} \text{ respectivement.}$$

Cette approche est généralement appélée **équité par conscience** (en anglais *fairness through awereness - FTA*) [13].
Pour évaluer ce critère sur un dataset donné, on peut citer deux métriques proposées dans la

littérature [4] :

• **La consistance** proposée par Zemer et al. (cité par Alessandro et al. (2022)), mesurant pour chaque individu l'écart de sa prédiction à celle de ses k plus proches voisins. Elle s'écrit :

$$Cons = 1 - \frac{1}{n}\Big(\sum_{i=1}^{n}\Big|\hat{y}_i - \frac{1}{k}\sum_{x_j \in kNN(x_i)}\hat{y}_j\Big|\Big) \text{ kNN}(x_i) \text{ les k plus proches voisins de } x_i \quad (2.1)$$

Les limites de cette métrique sont (i)- le choix de la distance pour la réalisation des $k - NN$ [2] , (ii)- le choix du nombre k optimal.

• La métrique de Berk et al. (2017)[3] qui mesure la moyenne des écarts de prédictions de toutes les paires d'individus de différents groupes de D, pondérés par l'inverse de leurs distances dans $\mathcal{X}$ (plus forte est la distance, moins important est l'écart). Cette métrique s'écrit :

$$\frac{1}{n_1 n_2}\sum_{(x_i,y_i)\in\{D=d_1\},(x_j,y_j)\in\{D=d_2\}} e^{-D(x_i,x_j)}|\hat{y}_i - \hat{y}_j| \quad (2.2)$$

où n_1 et n_2 sont les tailles d'échantillon dans les sous-groupes $\{D = d_1\}$ et $\{D = d_2\}$.
Cette approche souffre donc du problème du choix d'une distance convenable dans l'espace $\mathcal{X}$, étant donné qu'il peut y avoir des variables de plusieurs types (continues, binaires) et de différentes échelles. De ce fait, il n'est pas possible de proposer une métrique générale pour évaluer ce critère.

Une définition récente de l'équité tout aussi *aware* de l'existence de proxys est l'**absence de discrimination proxy** [3], proposée par Mathias L. et al (2024) [21]. Dans leur article, ils définissent un modèle exempt de discrimination par proxy comme celui dont les prédictions sont presque-sûrement identiques pour toute configuration comparable des variables en présence [4].La comparabilité ici renvoie à une situation où toute chose reste égale, sauf la relation de dépendance entre D et X. Ainsi, un modèle satisfait ce critère si (1) - il n'utilise pas directement D et (2) - l'information sur D portée probablement par X n'est pas non plus utilisée. En particulier, dans une configuration où X et D sont indépendants, les prédictions doivent être identiques. Ceci s'écrit dans le formalisme mathématique :

$$m(X, \mathbb{P}) = m(X, \mathbb{P}^{\perp}),$$
$$\mathbb{P}^{\perp}(Y|X, D) = \mathbb{P}(Y|X, Y), \mathbb{P}^{\perp}(Z) = \mathbb{P}(Z), \mathbb{P}^{\perp}(Y, X, D) = \mathbb{P}(Y|X, D)\mathbb{P}(X)\mathbb{P}(D).$$

Telle-que formulée, l'absence de discrimination par proxy d'un modèle ne paraît pas aisée à quantifier. Pour y parvenir, Mathias L. et al. (Juillet 2024) [20] définissent un ensemble théorique de modèles ne comportant pas de discrimination proxy. Ils proposent alors comme mesure pour un modèle donné la part de variabilité de ce dernier non expliquée par le modèle non proxy-discriminatoire le plus similaire à lui. Formellement, la métrique de proxy-discrimination pour un modèle m s'écrit :

$$PD\Big(m(X)\Big) = \frac{Var\Big(m(X) - \sum_{d\in\mathcal{D}} m(X,d)v_d^{\star}\Big)}{Var\Big(m(X)\Big)}, v^{\star} \in \mathcal{V} := \Big\{v \in [0,1]^{|\mathcal{D}|} : \sum_{d\in\mathcal{D}} v_d \leq 1\Big\} \quad (2.3)$$

2. Le *k-Nearest Neighbors* (**kNN**) est une méthode non paramétrique permettant d'identifier les k points les plus proches d'un point donné x_i en utilisant une métrique de distance. Les métriques courantes incluent la distance euclidienne et la distance de Manhattan, choisies selon la nature des données. Le choix de la métrique est essentiel, car il influence la pertinence des voisins sélectionnés, en particulier dans des espaces multidimensionnels.

3. Approximation, prédiction indirecte de la variable sensible par des features non-sensibles

4. Comparable dans la mesure où les distributions de probabilité des variables sont les mêmes, **sauf sur la structure de dépendance entre X et D**

• L'expression $\sum_{d\in\mathcal{D}} m(X,d)v_d^\star$ est celle du modèle exempt de discrimination proxy le plus proche du modèle m.
• On a : $0 \leq PD(m) \leq 1$, avec $PD(m) = 0$ si m ne comporte pas de discrimination proxy.
• Par convention, $PD(m) = 0$ si $Var(m(X)) = 0$.

Définitions fondées sur une structure causale sous-jacente

Contrairement aux définitions précédentes qui s'appuient sur des corrélations - pouvant être trompeuses - entre certaines features et D, cette approche repose sur une connaissance a priori des variables en présence. Cela se fait par un modèle causal qui décrit le mécanisme de génération des données, ce qui est crucial pour évaluer si une inégalité observée est d'une cause légitime ou pas. Dans ce cadre, un modèle satisfait l'**équité contrefactuelle** telle que définie par Kusner et al (2018) si la variable D ne cause - au sens du modèle causal sous-jacent - pour aucun individu ses prédictions.

Modèle causal

Un modèle causal au sens de M. Kusner et al. (2018) [18] est un triplet de vecteurs (U, V, F) où :

— U est un ensemble de variables non influencées, ou exogènes ;
— V comporte l'ensemble $\{V_1, ..., V_m\}$ des variables observables et endogènes ;
— F renvoie à l'ensemble de fonctions de liaison entre les variables exogènes et les endogènes. Pour tout i, $V_i = F_i(p_{a_i}, U_{a_i}), p_{a_i} \subset V | V_i, U_{a_i} \subset U$

Il est généralement représenté par un graphe acyclique orienté dans lequel les variables en présence sont les nœuds et leurs relations causales matérialisées par des flèches unidirectionnelles.

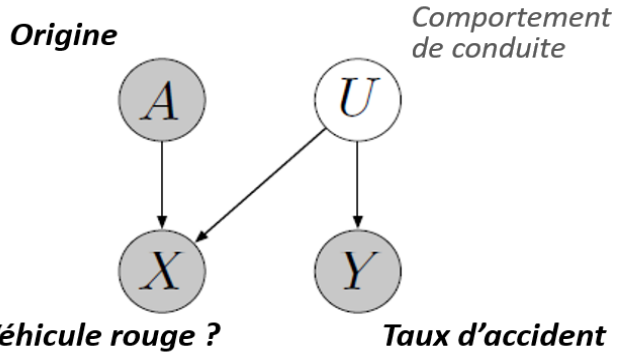

FIGURE 2.1 – Exemple de graphe acyclique orienté en assurance automobile

Interprétation : La préférence pour les voitures de couleur rouge est influencée par l'origine. De même, le comportement de conduite (aggressif) a un effet sur la préférence de cette couleur. Il influe aussi sur le niveau d'accident.

Pour évaluer ce critère, A. Charpentier (2024) propose de comparer chaque individu (x_i, d_i) à son "contrefactuel" - individu hypothétique appartenant à l'autre catégorie de D et ayant les caractérisques correspondantes $(\mathcal{T}(x_i), d_j)$ [5]. La correspondance se fait par le *transport optimal*

5. Par exemple, s'il s'agit de l'individu maximal (ayant des caractéristiques maximales X dans le sous-groupe $\{d_i\}$ alors son contrefactuel aura les valeurs maximales dans le sous-groupe $\{d_j\}$.

$(\mathcal{T})$ de la loi de X conditionnellement à $\{D = d_i\}$ à celle de X conditonnellemet à $\{D = d_j\}$; le principe est dit *mutatis-mutandis*.

$$\mathbb{E}\Big(\hat{Y}_{D=d_i}|X = x_i\Big) = \mathbb{E}\Big(\hat{Y}_{D=d_j}|X = \mathcal{T}(x_i)\Big).$$

Malgré l'avantage que présente l'équité contrefactuelle, en raison de son évaluation de l'impact causal de la variable sensible sur le modèle, elle n'est pas évidente à implémenter dans la pratique. En effet il n'est pas toujours aisé d'avoir un modèle causal dans une situation donnée (connaître toutes les variables en présence, et en avoir les données).

Principe	**Définition**	**Idée**	**Exemple d'évaluation**
Similitude	**Équité par ignorance** (Fairness through Unawareness, FTU)	Retirer la variable sensible du modèle	Pas d'utilisation explicite des variables sensibles.
	Équité par consistance (Fairness through Awareness, FTA)	Une distance faible dans les features ⇒ distance faible dans les prédictions	Consistance de Zemer et al. : comparer les prédictions des k plus proches voisins
Principes causaux	**Équité contrefactuelle**	Comparer chaque individu à son "contrefactuel" appartenant à une autre catégorie de D	Comparaison des prédictions entre l'individu réel et son contrefactuel (via le transport optimal)
	Absence de discrimination proxy	Les prédictions ne doivent pas utiliser implicitement ou explicitement la variable sensible D	Mesurer la part de variabilité non expliquée par un modèle non discriminatoire proche

TABLE 2.1 – Résumé des définitions de l'équité individuelle, des idées, et métriques d'évaluation

2.2.2 L'équité au niveau global ou équité de groupe

L'équité de groupe ou équité statistique, exige que les avantages ou les préjudices résultant d'un modèle soient répartis équitablement entre des groupes précis. Ces groupes sont définis à l'aide de la variable sensible / protégée. Ainsi, contrairement à l'approche individuelle, cette approche compare les groupes dans leur ensemble. La question centrale ici est de savoir si les groupes sont traités différemment ou non par le modèle.
La majorité des définitions de l'équité de groupe découlent de trois principes de base : l'indépendance, la séparation et la suffisance.

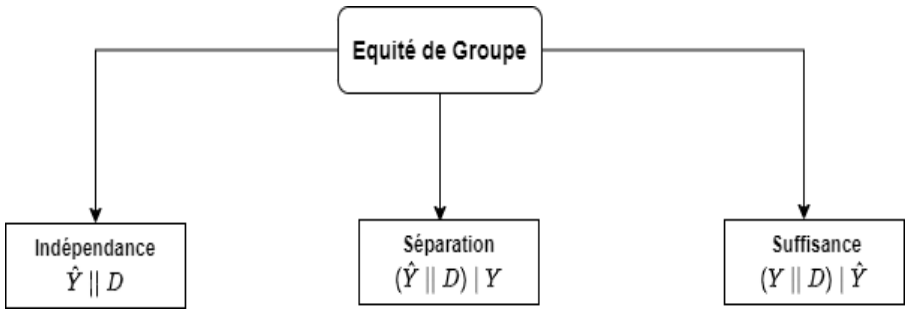

FIGURE 2.2 – Principes de l'équité de groupe

Le principe d'indépendance

Selon ce principe, l'équité d'un modèle repose sur l'indépendance entre ses prédictions et la variable sensible / protégée. On l'appelle aussi **parité démographique** ou **parité statistique**. Le formalisme statistique qui sous-tend cette définition change selon que le modèle est un classificateur ou un régresseur.

— Pour un modèle de régression

On distingue la *parité démographique forte* ou une indépendance vis-à-vis de la fonction de distribution : $\mathbb{P}(\hat{Y} \in \mathcal{A}|D = d_1) = \mathbb{P}(\hat{Y} \in \mathcal{A}|D = d_2)$ de la *parité démographique faible* : $\mathbb{E}(\hat{Y}|D = d_1) = \mathbb{E}(\hat{Y}|D = d_2)$.

— Pour un classificateur binaire

Les deux notions restent telles lorsqu'on s'intéresse aux scores $m(Z)$, mais se confondent lorsqu'on s'intéresse aux labels $\hat{Y}$ car les égalités suivantes sont équivalentes :

$$\begin{aligned} P(\hat{Y} = 1|D) &= P(\hat{Y} = 1) \\ P(m_\tau = 1|D) &= P(m_\tau = 1) \\ E(\hat{Y}|D) &= E(\hat{Y}) \end{aligned}$$

Prenons l'exemple d'un algorithme d'acceptation des sinistres, avec le genre [6] comme variable sensible. La parité démographique est vérifiée pour ce modèle si les chances d'accepter les réclamations sont les mêmes pour les hommes et les femmes.

6. Cette variable est considérée binaire dans la suite

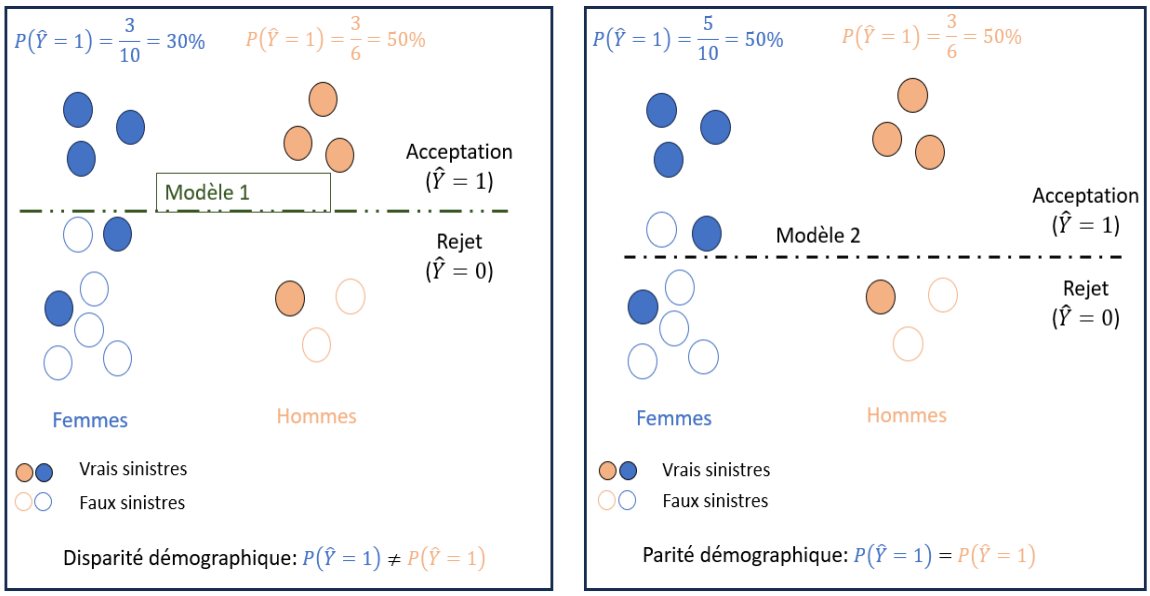

FIGURE 2.3 – Illustration de la parité démographique pour un classificateur binaire

Ce graphique montre à gauche un modèle où la parité démographique est vérifiée et à droite un où elle ne l'est pas

Dans certaines situations, on peut admettre des disparités de prédictions par rapport à des variables non-sensibles - ces disparités étant le reflet d'une forte hétérogenéité de la variable cible. La prime pure en assurance automobile par exemple, serait très dépendant du niveau de franchise contractuel ou du type de véhicule assuré. Dans ce cas, la définition adaptée serait la *parité démographique conditionnelle*. Sa signification est quasiment la même que précédemment, à ceci que les propriétés doivent être vérifiées sur des individus ayant certaines features identiques.
Ce principe d'équité de groupe ne prend pas en compte les dépendances, s'il en existe, entre Y et D. Cela aboutirait alors à un modèle qui ne réflèterait pas la réalité (lorsque les données y relative sont fiables). C'est pourquoi d'autres principes sont proposées ([5]).

Le principe de séparation

La séparation est vérifiée pour le modèle m si ses prédictions sont indépendantes de la variable sensible lorsque la variable cible Y est fixée. Autrement dit, lorsque toute dépendance entre D et les prédictions est expliquée par Y ; la population est *séparée* par Y et par rien d'autre. Ce principe se décline en plusieurs définitions selon le type de modèle et le degré d'exigence.

— Pour un modèle de régression

Les déclinaisons en séparation « forte » ou « faible » sont possibles comme pour l'indépendance et s'écrivent respectivement :

* $P(\hat{Y} \in \mathcal{B} \mid Y = y, D = d_1) = P(\hat{Y} \in \mathcal{B} \mid Y = y, D = d_2) \forall \mathcal{B} \subset \mathcal{Y}, y \in \mathcal{Y}$
* $\mathbb{E}(\hat{Y}|Y = y, D = d_1) = \mathbb{E}(\hat{Y}|Y = y, D = d_2), \forall y$

Dans la pratique cependant, leur évaluation est complexe, en raison de la continuité de Y (infinité ou grand nombre de y).
D'autres définitions sont proposées comme la ***similitude des erreurs*** par Zafar et al. (2019) cité par A. Charpentier ([5]) :

$$P(Y - \hat{Y} \in \mathcal{A} \mid D = d_1) = P(Y - \hat{Y} \in \mathcal{A} \mid D = d_2)$$

Cette contrainte peut être atténuée (et facilement évaluée) en ne considérant que l'indépendance de certains moments des résidus :

$$\mathbb{E}\Big(|Y-\hat{Y}|^a \mid D=d_1\Big) = \mathbb{E}\Big(|Y-\hat{Y}|^a \mid D=d_2\Big).$$

— Pour un modèle de classification binaire

Plusieurs critères d'équité se réfèrent au principe de séparation pour les classificateurs. La plupart d'entre d'eux indique la parité d'une ou de plusieurs métriques de performance lorsqu'on s'intéresse aux labels de prédiction. On a par exemple l'**égalité des opportunités** qui renvoie à la parité des taux de vrais positifs entre les différentes catégories de D :

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

Un autre critère est l'**égalité des chances** (Equal Odds en anglais), critère d'équité qui requiert en plus de l'égalité des taux de vrais positifs celle des taux de vrais négatifs.

$$P(\hat{Y}=1|Y=y, D=d_1) = (\hat{Y}=1|Y=y, D=d_2), \forall y \in \{0;1\}$$

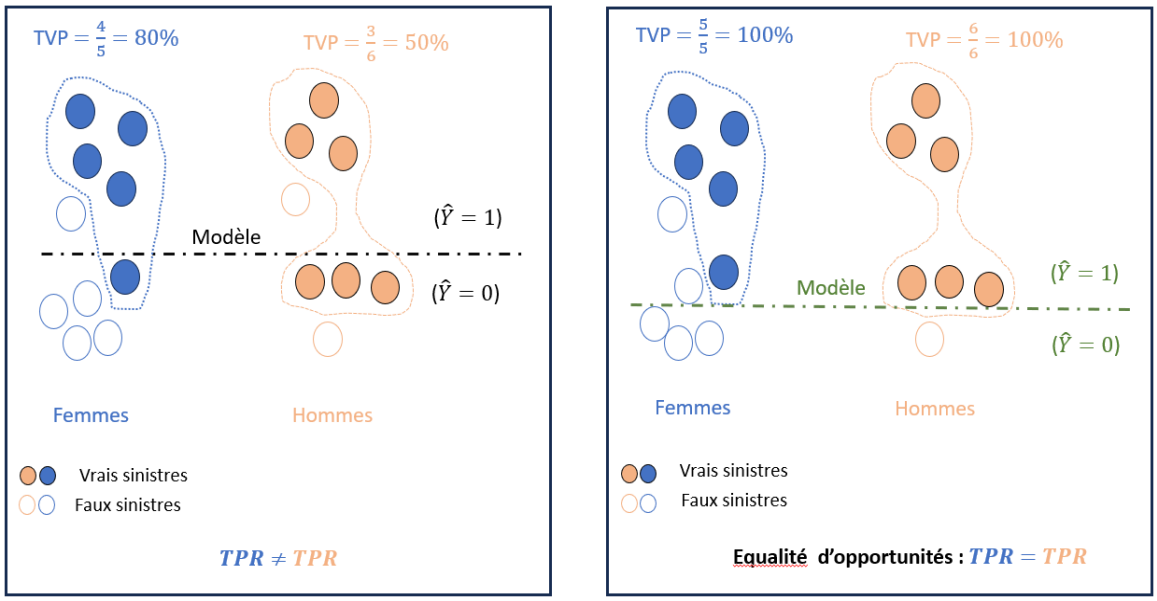

FIGURE 2.4 – Illustration de l'égalité d'opportunités pour un classificateur binaire

Les définitions précédentes s'étendent au cas des scores, en considérant l'égalité des fonctions de distribution en tout point. Les termes portent le qualificatif « forte », cela se comprend car une parité des distributions de scores est beaucoup plus exigeante que celle de quelques métriques.

- L' **égalité d'opportunités forte** est vérifiée si :

$$\mathbb{P}(m(Z) \in \mathbb{A}|D=d_1, Y=1) = \mathbb{P}(m(Z) \in \mathbb{A}|D=d_2, Y=1).$$

- L'**égalité des chances forte** ou **Equilibre des classes fort** est validée lorsque :

$$\mathbb{P}(m(Z) \in \mathbb{A}|D=d_1, Y=y) = \mathbb{P}(m(Z) \in \mathbb{A}|D=d_2, Y=y), \forall y \in \{0;1\}.$$

Lorsqu'on se restreint aux moyennes des scores, on parle d'**équilibre de classes faible**.

Plusieurs autres définitions sont proposées ou peuvent l'être, et reposent sur la parité de certaines métriques ou fonctions de métriques de performance. C'est le cas de la parité des AUC (Area Under Curve), celle des MCC (Coefficients de corrélation de Mattew), ou encore celle des courbes ROC (Receiver Operating Characterisctic). Mais en raison de leur complexité, elles ne sont pas présentées ici.
La séparation souffre du fait de nécessiter des données fiables sur la variable cible Y. De plus, si celle-ci est continue (cas d'un régresseur), l'évaluation pratique nécessite sa discrétisation qui peut être arbitraire.

Le principe de suffisance

C'est un principe qui adopte une perspective différente de celle de la séparation. La question est de savoir si toute dépendance entre la variable sensible et la variable d'intérêt ne repose que sur les prédictions. Dit autrement, il s'agit de vérifier que le modèle suffit (sans besoin de D) à expliquer les variations de Y. Ceci revient à dire que le niveau de calibration du modèle est identique dans les différents groupes de D. Formellement, la séparation est vérifiée lorsque : $Y \perp\!\!\!\perp D \mid \hat{Y}$.
Pour un modèle de régression,la suffisance s'écrirait :

$$\mathbb{E}\Big(Y \mid \hat{Y} = y, D = d_1\Big) = \mathbb{E}\Big(Y \mid \hat{Y} = y, D = d_2\Big), \forall y \in \mathcal{Y}$$

ou de manière stricte

$$\mathbb{P}\Big(Y \in \mathcal{A} \mid \hat{Y} = y, D = d_1\Big) = \mathbb{P}\Big(Y \in \mathcal{A} \mid \hat{Y} = y, D = d_2\Big), \forall y \in \mathcal{Y}, \mathcal{A} \subset \mathcal{Y}.$$

Comme exemple, dans le cadre d'une tarification automobile, la définition signifierait que le coût moyen de sinistres soit le même dans les différents groupes de D à niveau de prime payée y fixé.
Pour un modèle de classification, on parle de **Parité prédictive** lorsque la précision et la valeur prédictive négative (NPV [7] en anglais) sont égales dans les différents groupes de D - la target étant les labels. Et lorsque l'exigence d'équité porte sur le score, on parle de **parité de calibration** :

$$\mathbb{P}\Big(Y = 1 | m(Z) = t, D = d_1\Big) = \mathbb{P}\Big(Y = 1 | m(Z) = t, D = d_2\Big), \forall t \in [0; 1]$$

2.2.3 Les métriques d'évaluation de l'équité de groupe

Comme décrit plus haut, toutes les définitions d'équité de groupe s'écrivent sous forme d'indépendance probabiliste. Il n'est donc pas surprenant que les métriques d'évaluation dans la littérature soient calculées sur des principes similaires, selon qu'on évalue l'indépendance, la séparation ou la suffisance. Deux approches émergent :

— Comparer chacun des sous-groupes de D à l'ensemble, puis aggréger soit par un maximum ou une moyenne. Cette approche est facilement transposable au cas où la variable D est multiclasses ;

— Comparer les sous-groupes de D entre eux directement. Transposer cette approche de calcul au cas multiclasse consisterait à aggréger les résultats obtenus pour toutes les paires par une moyenne ou un maximum.

Ainsi, le plus important est la manière de comparer deux distributions de probabilité. Une différence doit cependant être faite entre les modèles de régression, et les classifications.

7. Negative Predictive Value

Pour les modèles de régression, Y continu

Les métriques visent à évaluer l'écart de distributions dans deux groupes de D. Ce qui n'est pas une problématique nouvelle en statistique. La comparaison des distributions peut être stricte ou moins rigoureuse. Comme métriques d'évaluation **stricte** de l'équité, on peut citer :

* **La distance de Wasserstein**. Elle évalue le « travail » minimal pour transformer une mesure de probabilité en une autre. Un régresseur vérifie le principe d'indépendance si et seulement si la distance de Wasserstein de ses prédictions dans les 2 groupes de D est nulle : $W_2(\mathbb{P}_{d_1}, \mathbb{P}_{d_2}) = 0$, ($(\mathbb{P}_{d_1}, \mathbb{P}_{d_2})$ mesures de probabilité des prédictions de m respectivement sur $\{D = d_1\}$ et $\{D = d_2\}$)[5]. La distance s'écrit :

$$W_p(\mu, \nu) = \left(inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{Y}\times\mathcal{Y}} |y_1 - y_2|^p d\pi(y_1, y_2) \right)^{\frac{1}{p}}, p > 0. \tag{2.4}$$

$\Pi(\mu, \nu)$, ensemble de mesures sur $\mathcal{Y} \times \mathcal{Y}$ dont les marginales sont μ et ν . Son implémentation dans le cas de comparaison de lois continues nécessite des hypothèses de discrétisation (pour le calcul des intégrales notamment).

* **La distance de Kolmogorov-Smirnov**. Elle est traditionnellement utilisée pour le test d'hypothèse qui porte le même nom et qui permet d'indiquer si deux échantillons suivent la même loi de probabilité ou pas ([26]) . Elle évalue l'écart maximal (normalisé) entre deux distributions empiriques. S'ils sont de tailles n m, avec comme fonctions de répartion empiriques $F_n(x)$ et $G_m(x)$, la distance s'écrit [25] :

$$D_{m,n} = Sup_{x\in\mathbb{R}} \left| F_n(x) - G_m(x) \right|$$

Son implémentation est beaucoup plus simple sans besoin d'hypothèses de discrétisation.

* **La divergence de Kullback-Liebler [29]**. C'est une autre manière de comparer deux distributions à travers une différence statistique ; ses auteurs l'appellent l' « information de discrimination entre deux populations ». Elle est beaucoup plus difficile à implémenter dans le cas des lois continues que dans le cas discret dont les expressions mathématiques sont respectivement :

$$KL(p||q) = \int_{\mathbb{R}^d} p(x) log \frac{p(x)}{q(x)} dx, KL(p||q) = \sum_i p(i) \frac{p(i)}{q(i)} \tag{2.5}$$

Pour une évaluation un peu plus **souple** de l'équité des modèles continus, les métriques utilisées reposent sur la comparaison de moyennes par des ratios ou des différences.

* **L'impact disparate (DI)**. Il s'agit simplement du ratio minimal des moyennes de la variable d'intérêt (prédictions) dans les 2 groupes de la variable protégée. Plus il est proche de 1, plus équitable est le modèle. Lorsque ce ratio est inférieur à un seuil τ, généralement 80%, le modèle est dit disparate. Dans le cas du principe d'indépendance, le ratio s'écrit :

$$DI(m) = min\left\{ \frac{\mathbb{E}\Big(m(Z) \mid D = d_1\Big)}{\mathbb{E}\Big(m(Z) \mid D = d_2\Big)}, \frac{\mathbb{E}\Big(m(Z) \mid D = d_2\Big)}{\mathbb{E}\Big(m(Z) \mid D = d_1\Big)} \right\} \tag{2.6}$$

Dans le cas du principe de séparation ($\hat{Y} \perp\!\!\!\perp D \mid Y$), la métrique peut être adaptée en calculant ce ratio minimum dans chacune des catégories de Y ($\{\mathcal{Y}_i, i \in \mathcal{I} \subset \mathbb{N}\}$), puis aggrégé par

moyenne pondérée de chacun des groupes [8] ou par minimum.

$$DI_{sep} = min_i\{DI(m \mid Y), Y \in \mathcal{Y}_i\}, \mathcal{Y}_i \subset \mathcal{Y}.$$

Ceci peut aussi être répliqué pour évaluer la suffisance, en adaptant bien évidemment à ce cas.

* **La Différence de parité statistique (SPD)**. Comme son nom l'indique, c'est une mesure d'évaluation de la parité statistique. Elle mesure plus exactement l'écart absolu entre les moyennes des prédictions du modèle dans les deux groupes de la variable D.

$$SPD(m) = \Big|\mathbb{E}(m(Z) \mid D = d_1) - \mathbb{E}(m(Z) \mid D = d_2)\Big| \tag{2.7}$$

Tout comme les métriques précédentes, celle-ci est également adaptable à l'évaluation de la séparation et de la suffisance, et aussi transposable au cas d'une variable sensible non binaire.

Pour les modèles de classification, Y binaire

Dans ce cas, l'équité peut être évaluée à 2 niveaux : soit au niveau des scores, soit au niveau des labels. Lorsqu'on s'intéresse aux **scores** (évaluation stricte), toutes les métriques précédentes peuvent être utilisées. En effet dans cette situation, la variable d'intérêt est continue (bien que restreinte sur $[0, 1]$) et les calculs précédents ne posent aucune difficulté.
Si l'intérêt est porté sur les labels, l'évaluation ne se fait pas exactement de la même manière. En plus de certaines métriques précédentes qui peuvent être adaptées (SPD, DI), d'autres métriques ont été proposées dans la littérature pour évaluer l'indépendance. On peut citer :

* **L'indice d'inéquité de groupe** - Group Unfairness Index (GUI) en anglais. Il permet de résoudre le problème de robustesse de l'estimation empirique de la différence de parité statistique lorsque les sous-groupes de D sont très différents (Siddiqi, 2012, cité par Charpentier 2024 [5]). L'indice est proposé par Gero Szepannek et al. (2021) [32] pour le credit scoring. Il est calculé de la manière suivante :

$$GUI(\hat{Y}, D) = \sum_{i \in \{0,1\}} \Big(\mathbb{P}(\hat{Y} = i \mid D = d_1) - \mathbb{P}(\hat{Y} = i \mid D = d_2)\Big) log \frac{\mathbb{P}(\hat{Y} = i \mid D = d_2)}{\mathbb{P}(\hat{Y} = i \mid D = d_1)} \tag{2.8}$$

Cette métrique est croissante avec l'absence d'équité.

* **L'information mutuelle** - Mutual Information (MI) en anglais. Elle permet de comparer la jointe de $(\hat{Y}, D)$ dans la configuration des données $\mathbb{P}(\hat{Y}, D)$ et dans la configuration d'indépendance $\mathbb{P}(\hat{Y})\mathbb{P}(D)$. Cette comparaison se fait en utilisant la distance de Kullback-Liebler dans le cas discret :

$$MI(\hat{Y}, D) = KL\Big(\mathbb{P}(\hat{Y}, D) \mid\mid \mathbb{P}(\hat{Y})\mathbb{P}(D)\Big) \tag{2.9}$$

Toutes les métriques citées précédemment sont utilisées pour l'évaluation du principe d'indépendance des modèles de classification. Pour les autres principes, de séparation et de suffisance, les métriques d'évaluation de l'équité s'appuient sur les métriques de perfomances correspondantes.

— Pour l'**égalité des opportunités** les taux de vrais positifs comparés dans les groupes de D par différence absolue ou par ratio ;

8. Problème, comment choisir ces groupes dans le cas continu ?

— Pour l'**égalité des chances** les taux de vrais positifs et de vrais négatifs sont comparés (par différences absolues ou par ratio minimum) puis agrégés par une moyenne ou un maximum ;

— Les précisions et valeurs prédictives négatives sont comparées de la même manière (ratio ou différence absolue) puis agrégées pour l'évaluation de la **parité prédictive**.

L'observation qui se dégage des sections précédentes est qu'il existe de multiples principes d'équité de groupe et donc de métriques d'évaluation y associées. Une question serait à ce stade s'il est possible pour un modèle de vérifier simultanément plusieurs critères d'équité de groupe.

Principe	**Classificateurs**	**Régressions (Target continue)**	**Exemples de Métriques**
Indépendance	— Parité démographique — Parité démographique conditionnelle	— Parité démographique forte — Parité démographique faible — Parité démographique conditionnelle (forte ou faible)	— Différence de parité statistique (SPD) — Distance de Wasserstein, Kolmogorov-Smirnov — Impact disparate (DI)
Séparation	— Égalité des chances — Égalité des opportunités	— Séparation forte et faible — Similitude des erreurs	— Différence ou ratio des rappels et/ou taux de vrais négatifs
Suffisance	— Parité prédictive — Parité de calibration	— Suffisance forte et faible	— Différence ou ratio des précisions et valeurs prédictives négatives (NPV) — Parité de calibration (égalité de la calibration)

TABLE 2.2 – Résumé des principes d'équité de groupe, et métriques associées.

2.2.4 Incompatibilité des critères d'équité de groupe

Plusieurs travaux récents ([23], [5], [4]) montrent qu'il est généralement impossible de satisfaire plusieurs critères d'équité de groupe en même temps, sauf dans des cas triviaux ou dégénérés. Cette incompatibilité découle de la nature différente des conditions imposées par chaque critère.

— Un modèle ne peut satisfaire l'**indépendance** et la **suffisance** pour une variable sensible D que si la variable d'intérêt est indépendante de la variable d'intérêt ($Y \perp\!\!\!\perp D$) ;

— Pour un modèle de classification, l'**indépendance** et la **séparation** ne sont atteignables simultanément que si $Y \perp\!\!\!\perp D$ ou si le modèle est complètement inutile ($m_\tau(Z) \perp\!\!\!\perp Y$) ;

— La **séparation** et la **suffisance** pour une même variable sensible D sont incompatibles sauf si (i)- les groupes de Y sont parfaitement équilibrés dans les groupes de D ($Y \perp\!\!\!\perp D$) ou si (ii)-le modèle est inutile ($m_\tau(Z) \perp\!\!\!\perp Y$).

Les conflits entre ces différents critères d'équité de groupe (la séparation, l'indépendance et la suffisance) reflètent les tensions inhérentes à la conception de modèles équitables. Ces résultats mettent en lumière la nécessité pour les praticiens de faire des compromis ou des choix entre différentes formes d'équité en fonction du contexte et des objectifs éthiques de l'application. Un exemple d'arbre de décision est proposé afin de faire le choix par rapport aux principes d'équité de groupe ([28]).

2.3 Evaluations des métriques sur données simulées

Dans l'optique d'une meilleure compréhension des métriques d'évaluation d'équité des modèles identifiés et proposés, nous avons trouvé intéressant de les évaluer. Les critères d'évaluation ont porté sur la sensibilité à la réalité mesurée et la capacité de classement des modèles. Ceci a été fait dans le but de sélectionner les meilleures métriques d'évaluation, en ayant en perspective la possibilité de les utiliser pour la comparaison des modèles ou l'évaluation des méthodes de mitigation de l'inéquité.
Pour y parvenir, un processus de simulation des données a été proposé avec des paramètres connus.

2.3.1 Processus de simulation des données

Deux simulateurs de données ont été employés, tous deux inspirés de l'article de MP Côté et al. (2024) [10]. Dans les deux cas, un jeu de données constitué de deux types de variables, une variable expliquée Y (continue puis binaire) et une ou plusieurs variable(s) explicative(s) X, est simulé avec des dépendances représentées sur le graphe causal ci-dessous. Comme dans la section précédente, D représente la variable sensible.

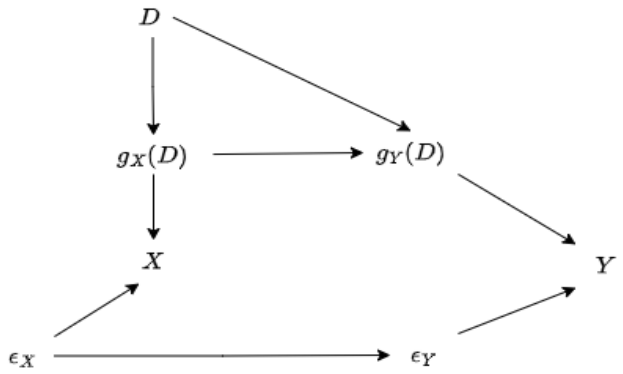

FIGURE 2.5 – Schéma causal du processus de simulation des données

La structure de dépendance est telle-que la variable sensible D cause simultanément X et Y au travers de $g_X(D)$ et de $g_Y(D)$ respectivement. Dans le même temps, ces deux variables (X et Y) ont des composantes indépendantes de D. La relation entre la variable expliquée et la variable explicative - reflet de la pertinence de celle-ci - est matérialisée par la causalité directe entre les compasantes de mêmes types.

2.3.2 Comparaison des modèles dans un cadre statique

Le premier processus de simulation des données utilisé est exactement celui de MP Côté et al. (2024), avec des paramètres figés. Le but est de savoir si les métriques proposées dans la section précédente permettent de rendre compte fidèlement d'une différence de niveau d'équité observée dans les données. Selon ce schéma :

— La variable sensible D est générée selon une loi de Bernouilli : $D \sim Ber(p)$;

— La variable explicative X est unidimensionnelle, $X = g_X(D) + \epsilon_X$ avec ϵ_X la partie de X indépendante de D telle-que : $\epsilon_X \sim \mathcal{N}(1, 9)$ et $g_X(D) = 10(D - 0.5)$;

— La variable expliquée Y est elle aussi résultat de 2 causes : $g_Y(D) = 15(D-0.5)+3g_X(D)$ la partie dépendante de D et ϵ_Y causée par le bruit ϵ_X indépendant de D.

$$(\epsilon_Y \mid \epsilon_X) = 100 + 3\epsilon_X + \mathcal{N}(0, 30)$$

Dans ce cadre, cinq modèles théoriques sont définis pour leurs niveaux de dépendance, directe ou indirecte avec la variable sensible D.

* **Modèle Best-estimate** : Ce modèle utilise toutes les informations disponibles, c'est-à-dire X et D, pour estimer la relation entre les variables explicative et expliquée. La prédiction est basée sur l'espérance conditionnelle complète :

$$m_{BE}(x, d) = \mathbb{E}[Y \mid X = x, D = d].$$

Cela signifie que le modèle est directement dépendant de D, ce qui entraîne une discrimination directe.

* **Modèle Unaware (ignorant)** :Ici, la variable sensible D est complètement ignorée dans la construction du modèle. On utilise uniquement la variable explicative X :

$$m_U(x) = \mathbb{E}[Y \mid X = x].$$

Bien que ce modèle ne discrimine pas directement, il peut toujours y avoir une discrimination indirecte si X est corrélé à D. De plus, ce modèle souffre d'un biais de variable omise ; la force de la dépendance entre X et Y capturée par le modèle, pourrait cacher l'effet direct de D sur Y

* **Modèle Aware** : Il est construit de manière à représenter la dépendance directe entre X et Y en excluant tout lien direct que peut avoir D sur Y. Cependant, le fait qu'il utilise l'information contenue dans X pour prédire Y ne supprime pas complètement les effets indirects de D sur ses prédictions via X :

$$m_A(x) = \mathbb{E}_D\mathbb{E}[Y \mid X = x, D] \quad .$$

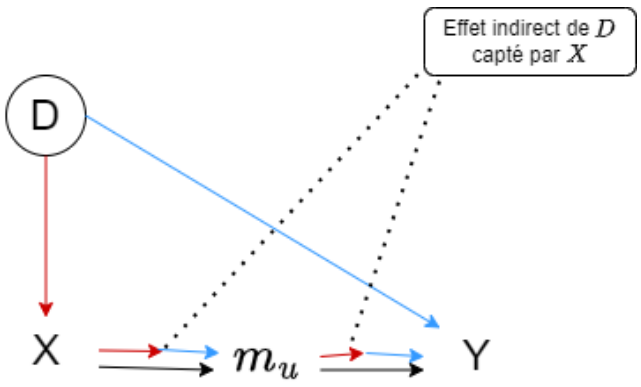

FIGURE 2.6 – Illustration du biais de variable omise
Lorsqu'on exclut D, la dépendance visible entre X et Y est artificiellement augmentée.

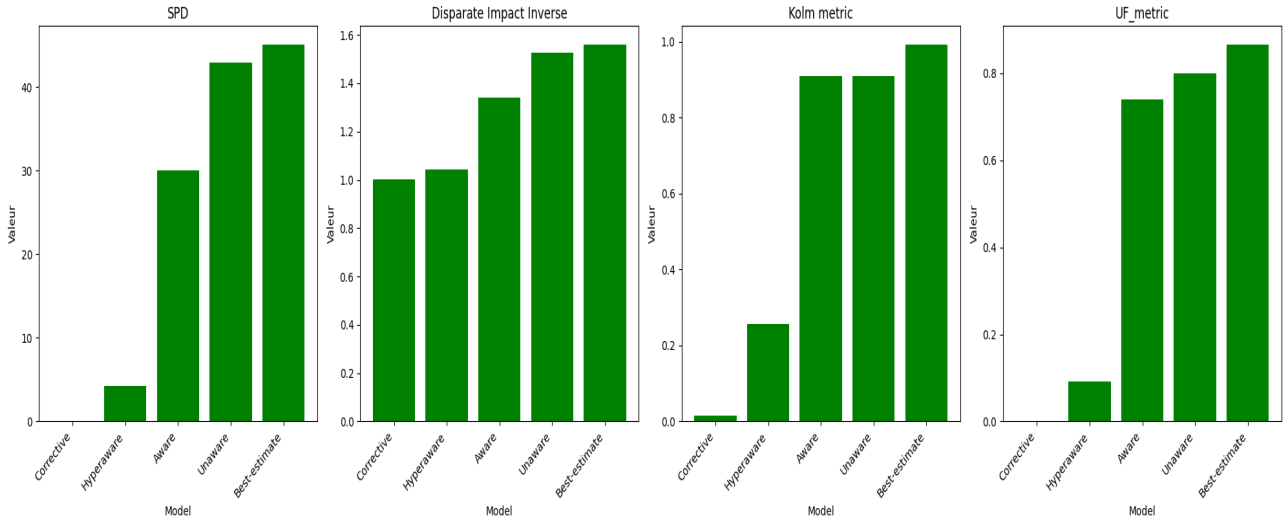

FIGURE 2.7 – Comparaison du niveau de parité démographique des modèles au travers de différentes métriques

* **Modèle Hyper-aware** : Celui-ci tente de réparer le défaut du précédent. Au lieu de prédire Y, la variable d'intérêt, il consiste à prédire la partie de Y indépendante de D, ϵ_Y. De ce fait, la dépendance indirecte entre D et Y n'est pas prise en compte. C'est ce modèle qui inspire certaines méthodes de mitigation des discriminations « ante-modélisation »(pre-processing en anglais) comme celle de Drago et al.(2021) ou même « in-modélisation »(en anglais in-processing).

$$m_H(x) = \mathbb{E}[\epsilon_Y \mid X = x].$$

Malgré cela, ce type de modèle utilise X, et de ce fait, pourrait être corrélée à D indirectement.

* **Modèle Correctif** : ce modèle est le plus avancé et vise à corriger explicitement toute forme de discrimination. Il ajuste la prédiction de manière à corriger les biais historiques. Toute dépendance entre la variable sensible et d'une part la variable expliquée, et d'autre part les variables explicatives n'est pas prise en compte.

$$m_C(\epsilon) = \mathbb{E}[\epsilon_Y \mid \epsilon_X = \epsilon]$$

où ϵ_X est la partie de X indépendante de D, garantissant ainsi qu'aucune information relative à D n'est utilisée dans le modèle.

Les paramètres sont choisis de manière à ce que tous ces modèles aient une valeur moyenne de prédiction identique. Lorsqu'on utilise le critère de **parité démographique**, l'ordre des modèles en matière d'équité est le suivant : $m_{BE}, m_U, m_A, m_H etm_C$ [10]. Cela se comprend assez aisément de par la manière dont sont construits ces modèles : le modèle Best-estimate utilise directement l'information sur D, tandis-que le correctif n'en utilise aucune. Cet ordre est visible avec quelques différences sur les métriques telles-que représentées ci-dessous.

Pour le critère de **Séparation** on a une autre relation d'ordre entre ces modèles. Comme le soulignent Olivier Côté et al ([10], p.24), les modèles best-estimate et correctif sont ceux qui violent le critère, tandis-que les trois autres présentent un comportement similaire. Dans ce cas encore, presque toutes les métriques le représentent assez fidèlement, sauf celles qui portent sur la parité des niveau d'erreur [9].

9. Erreur quadratique moyenne ou erreur absolue moyenne en pourcentage

Le critère de suffisance lui aussi fait ressortir une autre relation d'ordre entre les modèles. Rappelons qu'un modèle satisfait ce critère s'il suffit pour capter toute dépendance entre la variable d'intérêt et D. De ce fait, le modèle best-estimate est celui qui est le plus "équitable" au sens de ce critère. Le modèle correctif présente la plus grande violation de ce critère[10]. En plus des métriques fondées sur une parité des erreurs, celle qui repose sur une analyse de variance ne rendent pas compte de cette relation d'ordre. Ainsi, il est pertinent pour la comparaison de modèles en matière de séparation de ne considérer que la « Différence de parité statistique », l' « Impact disparate » et la distance de Kolmogorov comme métriques adaptées à la comparaison de modèles pour le critère de séparation.

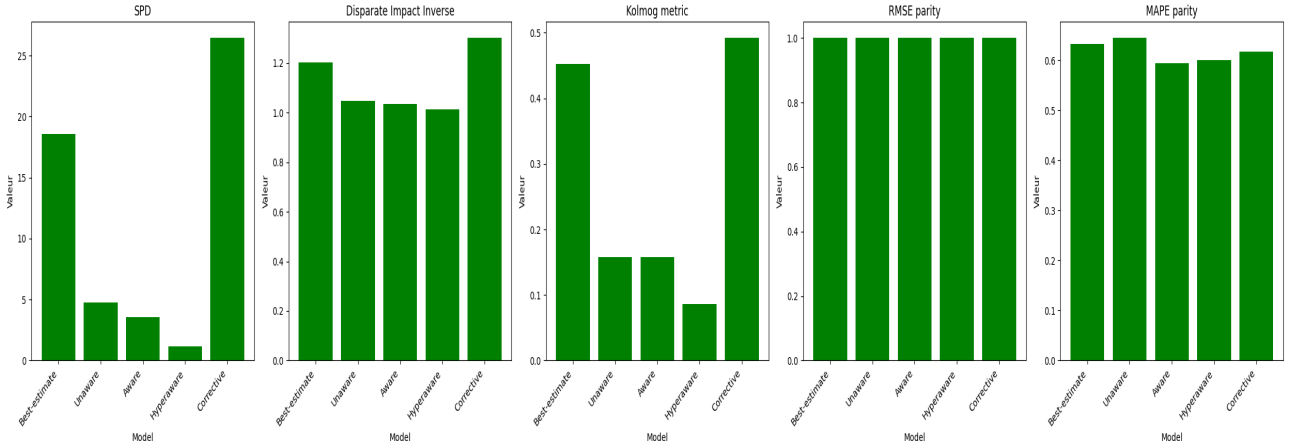

FIGURE 2.8 – Comparaison du niveau de Séparation des modèles au travers de différentes métriques

Les trois premières métriques réflètent bien l'ordre attendu ; les modèles Best-estimate et correctif sont les moins équitables. Mais, les deux derniers ne permettent pas de voir l'ordre entre modèles. Ils ne doivent donc pas être utilisés pour la classification des modèles en terme de séparation.

2.3.3 Simulations avec un paramètre dynamique : évaluation de la sensibilité

Pour évaluer la sensibilité des métriques au degré de dépendance entre dans les données, plusieurs paramètres sont introduits dans le processus de génération des données. Le paramètre principal est un coefficient k qui permet de moduler la dépendance entre la variable sensible D et les features du modèle. Aussi, la dimension des features non sensibles X est modifiée afin de se rapprocher d'un situation multi-dimensionnelle plus réaliste.On a un vecteur de variables explicatives (X_1, Z), où X_1 est dépendante de D, et Z ne l'est pas. La décomposition de X_1 se fait en une partie $g_{X_1}(D)$ et en une autre ϵ_{X_1}.

$$X = (X_1, Z), \text{ où } : X_1 = g_{X_1}(D) + \epsilon_{X_1} = kD + \mathcal{N}(\mu, \sigma^2), \text{ avec } k, \mu, \sigma \in \mathbb{R} \text{ et } Z \sim E(\lambda)$$

Par ailleurs, la variable expliquée Y est mise en relation avec toutes ces variables exogènes par des liaisons linéaires pour la plupart - avec tout de même une partie non linéaire - de manière à rester cohérent avec la décomposition précédente en 2 composantes , une dépendante de D et l'autre indépendante.

$$Y = g_Y(D) + \epsilon_Y(D) = \alpha_1 X_1 + e^{\alpha_2 X_1} + \alpha_3 Z + \alpha_4 D + (1 + D)\sigma_\epsilon \text{ avec } \sigma_\epsilon \sim \mathcal{N}(0, \epsilon^2)$$

$$Y = \left(\alpha_1 g_{X_1}(D) + e^{\alpha_2 (g_{X_1}(D) + \epsilon_{X_1})} + (\alpha_4 + \sigma_\epsilon)D\right) + \left(\alpha_1 \epsilon_{X_1} + \alpha_3 Z + \sigma_\epsilon\right)$$

Dans cette configuration, le nombre de paramètres semble important. Mais, il est clair que les plus influents sur le niveau de dépendance entre D et x_1 et entre D et Y sont k et α_2. Afin de faciliter les analyses la valeur α_2 est choisie de sorte que l'effet de k soit prédominant. Les valeurs des paramètres (en dehors de k) sont résumés dans le tableau joint en annexe 5.3.
Quatre modèles (sur les cinq présentés précédemment - en raison de leur simplicité) sont simulés et permettent d'évaluer la sensibilité des métriques.

— Le modèle best-estimate :

$$m_{BE}(x,d) = \mathbb{E}(Y|(X_1,Z) = (x_1,z), D = d) = \alpha_1 x_1 + e^{\alpha_2 x_1} + \alpha_3 * z + \alpha_4 * d;$$

— Le modèle ignorant ou unaware : $m_{UA}(x) = UA = \mathbb{E}(Y|(X_1,Z) = (x_1,d)) = \alpha_1 x_1 + e^{\alpha_2 x_1} + \alpha_3 z + \alpha_4 \mathbb{E}(D|(X_1 = x_1, Z = z))$ où on a :

$$E(D|(X_1,Z) = P(D = 1|X_1,Z) = \frac{f_{(X_1|D)}(x_1|D=1)P(D=1)}{f_{(X_1|D)}(x_1|D=1)P(D=1) + f_{(X_1|D)}(x_1|D=0)P(D=0)}.$$

— Le modèle aware : $m_A(x) = m_{BE}(x,d_1)\mathbb{P}(D = d_1) + m_{BE}(x,d_2)\mathbb{P}(D = d_2)$;

— Le modèle hyper-aware : $m_H(x) = \mathbb{E}(\epsilon_Y \mid X = x) = \mathbb{E}(\alpha_1 \epsilon_{X_1} + \alpha_3 Z + \sigma_\epsilon \mid X = x)$

Le principe d'équité de groupe évalué est l'indépendance. En effet, comme évoqué précédemment, les principes d'équité de groupe reposent tous sur des conditions d'indépendance, conditionnelles ou pas. La dépendance entre les prédictions et D est mesurée par le coefficient de corrélation linéaire. Cet indicateur est prévu pour mesurer la dépendance entre deux variables continues. Dans notre cas, D est une variable bianire (codée en 0 et 1) et peut être interprétée comme probabilité d'appartenir à l'un des groupes de D. Les analyses portent sur quatre métriques à savoir la distance de Kolmogorov, le ratio minimal des moyennes (Impact disparate), la différence absolue des moyennes (ou différence de parité statistique) et la part de variance expliquée. Les trois premières sont évaluées différemment, soit par des comparaisons entre groupes, soit par des comparaisons avec l'ensemble.

On observe (pour les quatre modèles) que toutes les métriques évoluent dans le sens souhaitable avec le niveau de dépendance absolu (décroissant pour le ratio minimal des moyennes et croissant pour les autres). De plus, la comparaison entre groupes semble plus sensible aux fluctuations de dépendance que la comparaison avec l'ensemble.

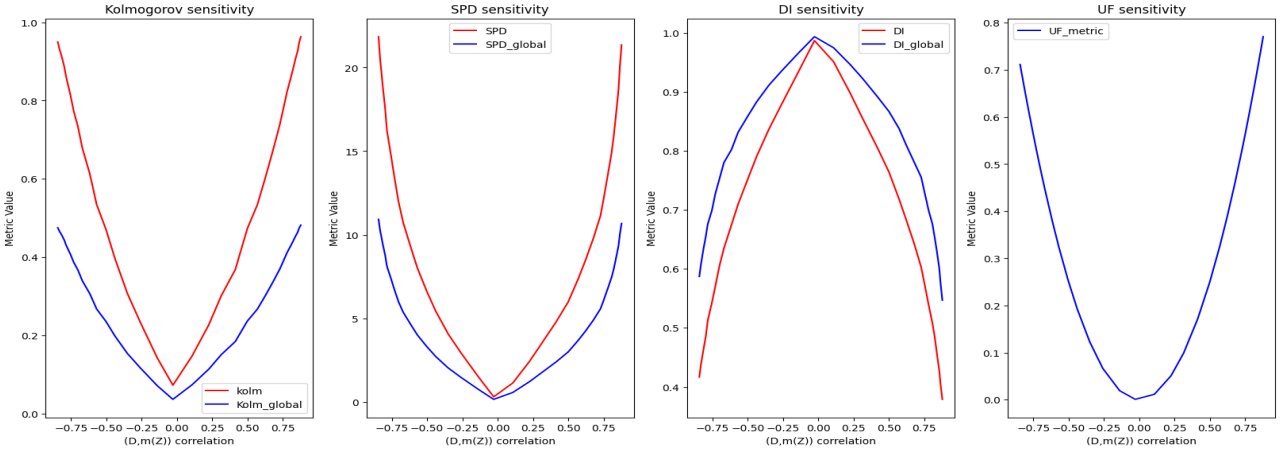

FIGURE 2.9 – Sensibilité des métriques au niveau de('in) dépendance
En abscisses, le degré de dépendance du modèle avec D, et en ordonnée la valeur de la métrique

Il est possible d'utiliser l'Analyse de Variance (ANOVA) pour comprendre la dépendance entre D et les prédictions. Dans cas, le lien est mesuré au travers de la part de variance de

la variable continue expliquée par D. Le résultat obtenu est globalement le même que précédemment. On note toutefois des irrégularités dans l'évolution des trois premières métriques. Ces irrégularités peuvent être attribuables à la construction des trois métriques en question. Un changement de la variance expliquée peut ne pas forcément induire un changement de ces métriques.

Pour la classication, les simulations précédentes sont reprises mais avec une variable d'intérêt une binaire. Cela se fait à travers la transformation par la fonction sigmoide [10] afin de ramener les valeurs en scores compris dans l'intervalle $[0; 1]$. Le résultat est alors un score. La variable d'intérêt est modélisée par une variable de loi Bernoulli avec le score comme paramètre. Les modèles quant à eux sont construits comme des indicatrices de scores avec des seuils choisis convenablement de manière à équilibrer au mieux les données. En faisant varier le paramètre k on calcule les valeurs des métriques et les mesures de dépendance entre score et variable sensible. Les observations sont alors similaires au cas des régressions pour toutes les métriques, y compris l'indice d'inéquité de groupe et l'information mutuelle.

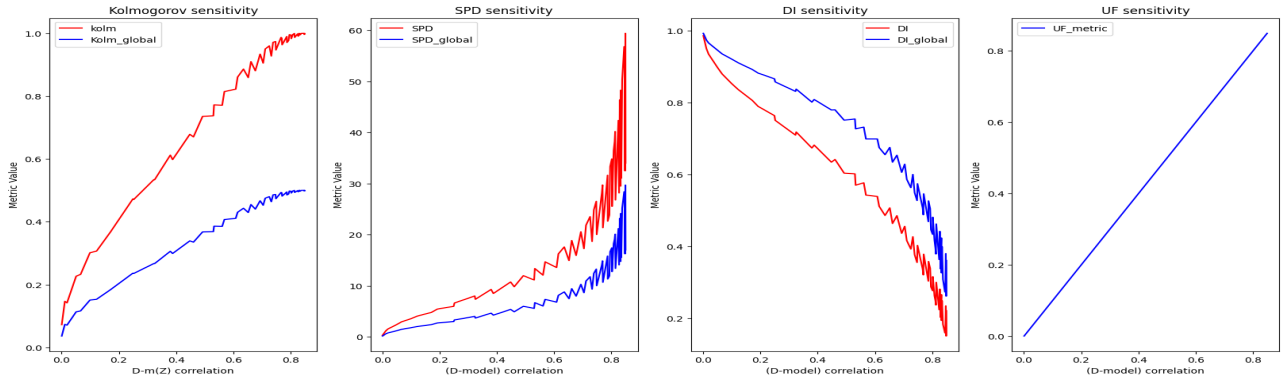

FIGURE 2.10 – Sensibilité des métriques au niveau de('in) dépendance -

En abscisses, le degré de dépendance du modèle avec D représentée par la part de variance expliquée, et en ordonnée la valeur de la métrique

10. $f(x) = \frac{1}{1+e^{-\lambda x}}$

Chapitre 3

Méthodes de mitigation de l'inéquité des modèles

Après avoir défini l'équité des modèles et proposé des métriques permettant de l'évaluer, la question est de savoir comment réduirer ou corriger un modèle qui est en marge d'un principe d'équité. L'intervention peut se faire à trois niveaux :

- **Sur les données** il s'agit de corriger les biais ou sources de discriminations du modèle à la source - on parle de méthodes pre-processing ;
- **Sur l'algorithme** en ajoutant des contraintes d'équité au problème d'optimisation du modèle ;
- **Sur les prédictions** en les transformant de manière à obtenir la condition d'équité satisfaite, on parle généralement de méthodes post-processing .

Le but de ce chapitre est de faire un état des lieux des méthodes existantes dans la littérature, afin d'en avoir une bonne compréhension. Quelques illustrations sont faites sur des données simulées et les limites de chacune, s'il en existe, sont présentées. Les notations utilisées restent celles du chapitre précédent.

3.1 Méthodes de mitigation Pré-modélisation

Les techniques de « pre-processing »consistent à transformer les données d'entrainement du modèle étudié de manière à satisfaire le critère d'équité souhaité. Une manière de le faire est de simplement retirer la variable sensible des données d'entrainement. D'autres méthodes vont au-delà : en plus de l'exclusion explicite de la variable sensible des données, elles font une exclusion de toute présence implicite de celle-ci dans les variables non sensibles.

3.1.1 L'exclusion de la variable sensible ou de ses proxys

Une manière naïve pour éviter des discriminations d'un modèle sur des groupes de la variable sensible D est de l'exclure de cette dernière lors de l'entrainement. Comme indiqué précédemment, c'est une approche qui est en accord avec la Gender Directive, en tarification assurantielle - la variable sensible étant le genre. Le modèle obtenu correspond à l'une des catégories de modèles « équitables »proposées par Olivier Côté et al ([10]), les modèles unaware(ignorants).

Lorsque certaines variables non-sensibles contiennent de l'information sur D - ce qui est très souvent le cas - cette technique n'est pas la plus efficace. Les discriminations directes cèdent la place à des traitements disparates indirects, du fait du bais de variables omises. Par exemple en assurance automobile, le type de véhicule assuré peut être un proxy du genre, même si cette variable n'est pas utilisée

Une alternative intuitive immédiate est alors de supprimer non seulement D, mais toutes les composantes X_j de X comportant de l'information significative sur D. L'application de cette approche conduira à une baisse de performance du modèle, ce qui nécessitera de choisir le niveau de significativité de dépendance entre D et les X_i de manière à arbitrer entre ces deux caractéristiques du modèle. D'autres méthodes considèrent toutes les composantes de X, mais en les transformant.

3.1.2 L'orthogonalisation des features non sensibles X

L'orthogonalisation des features vise à décorreler les variables explicatives non sensibles à la variable sensible. Par exemple, si D est la variable sensible (comme le genre) et $X = (X_1, ..., X_j, ..., X_d)$ le vecteur de variables explicatives non sensibles, la méthode consiste à transformer chacune des variables X_j en une nouvelle variable $X_j^{\perp}$ orthogonale à D. L'objectif recherché ici est d'obtenir un modèle qui satisfait le principe d'indépendance.

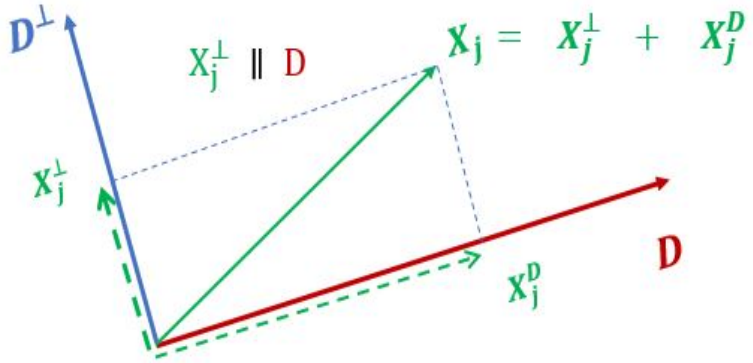

FIGURE 3.1 – Idée de l'orthogonalisation

Résoudre ce problème revient à trouver $X_j^{\perp}$, la projection orthogonale à D de chacune des features X_j. Dans le cas général de plusieurs (k) variables sensibles en présence, notons $D = (D_1, ..., D_k)$ la matrice $n \times k$ de leurs observations. La projection orthogonale à D est définie par la matrice $\Pi_{D^{\perp}}$, de sorte que pour une variable X_j le projeté orthogonal soit donné par $\Pi_{D^{\perp}} X_j$. Cette matrice est définie comme suit :

$$\Pi_{D^{\perp}} = \mathbf{I} - D(D^{\top}D)^{-1}D^{\top}$$

En pratique, le projeté orthogonal $X_j^{\perp}$ peut être vu comme le résidu de la régression $\hat{X}_j$ de X_j sur D obtenu par la méthode de minimisation des moindre carrés ordinaires.

$$X_j^{\perp} = X_j - \mathbb{E}[X_j \mid D] = X_j - X_j^D$$

Le modèle corrigé est obtenu par entrainement de la même structure de modèle sur les projetés $X^{\perp}$. Bien que simple à réaliser, cette méthode présente quelques inconvénients :

— **La signification des projections obtenues**

Dans le cas de l'existence d'une corrélation - même faible - avec D, les valeurs de chaque variable non sensible sont modifiées. On pourrait passer de valeurs strictement positives comme l'âge des assurés en assurance automobile, à des valeurs relatives (positives ou non). Pour y remédier, il est possible de contrôler la part d'information de la variable initiale qui sera projetée. Cela peut être fait au moyen d'un paramètre α, reflétant la proportion d'information initiale conservée. La nouvelle variable transformée est :

$$X_j^* = (1-\alpha)X_j + \alpha X_j^{\perp}, \alpha \in [0,1].$$

— **Les modèles non linéaires**

Même si $X_j^{\perp}$ et D sont décorrelées, cela ne signifie pas qu'elles sont indépendantes. De plus, l'indépendance $X_j \perp\!\!\!\perp D$ est vérifiée, cela n'implique pas forcément l'indépendance avec les modèles obtenus avec X_j (on n'a pas forcément $\Phi(X_j) \perp\!\!\!\perp D$). Ceci peut être le cas notamment des modèles non linéaires (modèles d'arbres, processus gaussien, réseaux neuronaux). Un moyen de contourner ce défaut serait l'utilisation des résidus d'un modèle non linéaire de prédiction des variables explicatives X_j par D. Mais, le problème d'interprétabilité des variables transformées restera non résolu. C'est entre autres pourquoi d'autres méthodes sont proposées.

3.1.3 La méthode du transport optimal

L'idée de cette approche est la même que celle de l'orthogonalisation : transformer les caractéristiques non sensibles X en données indépendantes de D la variable sensible. Mais, la méthode ici est différente. Comme décrit par Mathias Lindholm et al (2024) ([21]), elle consiste à transporter les caractéristiques X dans un espace probabilisé où ils sont indépendants de D. Ainsi, la procédure de transformation se fait en deux étapes.

— **Choisir la distribution d'indépendance cible**

Il s'agit de la distribution de probabilité des caractéristiques X indépendantes de D. Formellement, on part d'une distribution de X conditionnellement à D qui dépend de la valeur prise par cette dernière :

$$X_d := X \mid_{D=d} \ \sim \ F_d.$$

La distribution cible est quand à elle indépendante des valeurs prises par D. Les features transformées seraient alors telles-que :

$$X^* \mid_{D=d} \ \sim \ F^*.$$

où F^* est une loi de probabilité indépendante de D définie sur le même espace que X.
Dans la littérature, il n'existe pas de méthode universelle pour le choix de cette distribution indépendante.

— **Transformer chacune des observations de l'espace dépendant de D à l'indépendant**

Il s'agit de construire une fonction qui transforme chacune des données observées x (dans l'espace probabilisé où X et D sont dépendants) en des données indépendantes de D, x^* (dans l'espace de probabilité dont la distribution est F^*) avec le minimum d'effort ou de distance entre x et x^*. Cela nécessite de choisir une mesure de distance ou d'effort dans l'espace des features X. Le résultat est une fonction $T_d : X_d \mapsto X^*$.

La fonction peut être calculée explicitement dans le cas unidimensionnel d'une variable X continue. En effet, la fonction $f : x \mapsto f(x) = (F^*)^{-1}(F_d(x))$ permet de réaliser le transport de manière optimale avec la distance euclidienne comme fonction de coût ([21]). Cependant, dans les cas multidimensionnels ou en présence de variable catégorielle, la résolution est beaucoup plus complexe.

Contrairement à l'orthogonalisation, cette méthode ne change pas la signification des variables transformées. En effet, le fait de choisir une loi de probabilité F^* définie sur le même espace que les données initiales X le garantit. Le transport optimal peut donc être vu comme une transformation *locale* de l'espace des variables explicatives.

Le problème initial pourrait être résolu directement. Pour rappel, le but est de transformer les variables explicatives X en variables Z indépendantes de D et le plus similaire possible aux

données initiales. Cette formulation ressemble au problème abordé par Delbaen et Majumbdar ([11]), qui montrent l'existence mathématique de la solution du programme :

$$Z^* = argmin_{Z \perp\!\!\!\perp D} \|X - Z\|_2$$

Malgré l'existence théorique de solution, les algorithmes concrets permettant de construire cette solution n'ont pas encore été construits dans la littérature.

3.2 Méthodes pendant la modélisation (in-processing)

Les méthodes pendant la modélisation modifient directement l'algorithme d'apprentissage afin que les résultats des modèles soient plus équitables au sens d'un critère choisi. Elles consistent généralement à ajouter une contrainte d'équité au programmes d'optimisation des modèles de machine learning classiques. Ces techniques sont plus complexes à mettre en œuvre, mais sont plus générales dans la mesure où elles peuvent être utilisées pour autre principe d'équité de groupe en dehors de l'indépendance.

3.2.1 Méthodes de réduction

Les méthodes de réduction offrent une approche flexible et générale pour introduire des contraintes d'équité dans les modèles de machine learning. L'idée principale consiste à reformuler le problème initial sous contrainte en une série de sous-problèmes standards, tels que des tâches de classification ou de régression pondérées, qui peuvent être résolues avec des algorithmes existants. En utilisant des multiplicateurs de Lagrange, ces méthodes traduisent les violations des contraintes en pénalités dynamiques, intégrées à la fonction de coût. Chaque itération de l'algorithme ajuste les pondérations des échantillons et les pénalités associées, réduisant progressivement les violations des contraintes jusqu'à atteindre un équilibre entre précision et équité. Le terme "réduction" reflète cette transformation progressive, où un problème complexe est réduit à une séquence de tâches plus simples.

Méthodes pour les classificateurs

Les méthodes de réduction en classification permettent d'introduire des contraintes d'équité dans les modèles en reformulant le problème en une série de tâches. L'idée de base est d'ajouter au programme d'optimisation du modèle une contrainte supplémentaire d'équité. Cela induit automatiquement une baisse de performance. Comme indiqué par A.Agarwal et al. (2018) [2], il est alors préférable de construire des classificateurs aléatoires pour un meilleur compromis équité-précision.

De tels modèles consistent à selectionner un modèle $\tilde{m}$ de l'espace $\mathcal{M}$ tous les classificateurs possibles (avec les données disponibles) et de l'utiliser pour faire les prédictions. Les modèles sont distribués dans $\mathcal{M}$ selon un loi de probabilité notée Q . Le problème d'optimisation sous contrainte d'équité s'écrit à la base sous la forme suivante :

$$min \sum_{\tilde{m} \in \mathcal{M}} Q(\tilde{m}) \times err(\tilde{m}) \text{ s.c. } M\mu(Q) \leq c,$$

où :

- M est une matrice capturant toutes les équations relatives à la contrainte d'équité.
- $\mu(Q)$ est un vecteur de moments conditionnels définis par :

$$\mu_j(Q) = \mathbb{E}_{(X,D,Y)\sim\mathcal{P}} \left[g_j(X, D, Y, m(X)) \mid E_j \right],$$

où g_j est une fonction représentant les moments pertinents (par exemple, pour la parité démographique, $g_j(X, D, Y, m(X)) = m(X)$) et E_j décrit des sous-groupes (par exemple $\{D = d\}$).

— c est un vecteur représentant les seuils tolérés pour les violations des contraintes.

Ce programme peut être réécrit à l'aide du multiplicateur de Lagrange λ vectoriel. La fonction de coût principale est définie comme $L(Q, \lambda)$, et le problème de d'optimisation précédent se réécrit :

$$min_{Q\in\Delta} \ \ max_{\lambda\in\mathbb{R}^{|\mathcal{D}\times\{-,+\}|}} \ L(Q, \lambda)$$

où :

— $L(Q, \lambda) = \mathbb{E}_{(X,Y,A)\sim\mathcal{D}}\left[\text{err}(Q)\right] + \lambda^\top \left(M\mu(Q) - c\right)$

— Δ représente l'ensemble des distributions possibles sur les classificateurs ;

— λ est un vecteur de multiplicateurs de Lagrange pondérant l'importance des contraintes dans l'optimisation.

L'idée du programme est de trouver un équilibre entre deux "joueurs" ayant des objectifs opposés. Le premier (Q) recherche à minimiser l'erreur de classification tout en respectant les contraintes d'équité. Le second quant à lui (λ) recherche une violation maximale de la contrainte d'équité. Son but dans le programme est de "forcer" le modèle à respecter les contraintes en attribuant un coût proportionnel aux violations de l'équité.

Un exemple est la **parité démographique**, où l'objectif est d'assurer que les taux de sélection (pour un modèle d'acceptation de sinistres) soient identiques entre les groupes définis par la variable sensible D. Les contraintes associées sont formulées comme suit :

$$\mu_d(Q) = \mathbb{E}[m(X) \mid D = d] - \mathbb{E}[m(X)] \leq 0.$$

Ces contraintes sont intégrées à la fonction de coût via le terme $\lambda^\top(M\mu(Q) - c)$. Les multiplicateurs λ ajustent dynamiquement l'importance des contraintes en donnant plus de poids aux violations importantes. [1].

Méthodes pour les régressions

Pour les régressions, le principe de base est le même. Puisque les prédictions $m(X)$ sont continues, les contraintes d'équité sont adaptées pour capturer des distributions continues. Le problème d'optimisation contraint de base se formule comme suit :

$$\min_{m\in\mathcal{F}} \mathbb{E}\left[\ell(Y, m(X))\right], \quad \text{sous réserve de} \quad |\ \mathbb{P}[m(X) \leq z \mid D = d] - \mathbb{P}[m(X) \leq z]\ | \leq \zeta_d,\ \forall d \in \mathcal{D}, z \in [0, 1]$$

où :

— $\ell(Y, m(X))$ est une fonction de coût Lipschitzienne, par exemple la perte quadratique $(Y - m(X))^2$,

— ζ_d est un seuil spécifié pour chaque groupe protégé $d \in D$, indiquant la perte maximale acceptable pour ce groupe.

Résoudre ce programme n'est pas aisé en raison de la nature continue des prédictions et du nombre potentiellement infini de contraintes induites par les distributions conditionnelles. Pour surmonter cet obstacle, Alekh Agarwal et al. (2019) [1] proposent une méthode inspirée de la réduction pour les classificateurs, basée sur une **discrétisation de l'espace des prédictions**.

1. Pour plus de détails sur l'algorithme, voir [2], P4

Principe de discrétisation : La discrétisation consiste à approximer l'espace continu des prédictions $m(X)$ par une grille discrète de valeurs possibles. Soit $Z = \{z_1, z_2, \ldots, z_N\}$ une grille uniforme de points couvrant l'intervalle des prédictions possibles, avec une granularité définie par $\alpha = 1/N$. La fonction de coût discrétisée est alors définie comme :

$$\ell_\alpha(y, u) = \ell\left(\bar{y}, \lfloor u \rfloor_\alpha + \frac{\alpha}{2}\right),$$

où $\bar{y}$ est la valeur dans une couverture discrète de y, et $\lfloor u \rfloor_\alpha$ est la valeur de u arrondie à l'entier multiple inférieur de α.

Réduction à un problème de classification pondéré : En utilisant cette discrétisation, le problème de régression sous contrainte peut être transformé en un problème de classification pondéré. Concrètement :

1. Les prédictions $m(X)$ sont remplacées par leurs versions discrètes $\bar{m}(X)$, ce qui permet de ramener les contraintes continues à un ensemble fini de contraintes indexées par les valeurs $z \in Z$.
2. Les contraintes d'équité sont reformulées en termes de proportion d'occurrences sur la grille discrète Z, réduisant ainsi le problème à une tâche de classification avec des seuils dépendant de la grille.

Le problème d'optimisation devient alors :

$$\min_{Q \in \Delta} \text{loss}_\alpha(Q), \quad \text{sous réserve de} \quad |\mathbb{P}[\bar{m}(X) \geq z \mid D = d] - \mathbb{P}[\bar{m}(X) \geq z]| \leq \epsilon, \, \forall d, z,$$

où $\text{loss}_\alpha(Q)$ est la perte discrétisée sur l'ensemble de prédictions randomisées Q.

Avantages de la méthode : Cette discrétisation permet de ramener le problème continu à un cadre de classification standard, pour lequel des algorithmes bien établis existent. Grâce à la réduction, il devient possible d'appliquer des solveurs de classification pondérée tout en garantissant un respect des contraintes d'équité au sein de chaque groupe protégé.

Les méthodes de réduction offrent une grande flexibilité en permettant l'utilisation de modèles existants comme "boîtes noires". En adaptant dynamiquement les pondérations ou les contraintes, elles permettent de trouver un compromis entre précision et équité. Cependant, les contraintes d'équité prises en compte dans les implémentations actuelles sont uniquement l'indépendance et la séparation.

3.2.2 Le mitigateur adversarial

Le mitigateur adversarial s'inspire des réseaux antagonistes génératifs (GANs) pour réduire les biais présents dans les modèles d'apprentissage supervisé. Son principe repose sur l'entraînement simultané de deux réseaux neuronaux ayant des objectifs opposés :

— Un *prédicteur* chargé d'apprendre à prédire la variable cible Y à partir des données X

— Un *adversaire* qui cherche à prédire une variable protégée D à partir des prédictions du prédicteur ($\hat{Y}$)

L'objectif global est d'amener le prédicteur à minimiser son erreur de prédiction tout en maximisant l'incertitude de l'adversaire sur D. Cela permet de réduire les informations indésirables sur D présentes dans $\hat{Y}$, ce qui favorise une meilleure équité dans le modèle. L'idée d'utiliser cette structure est proposée par B. Zhang et al. (2018)[33].

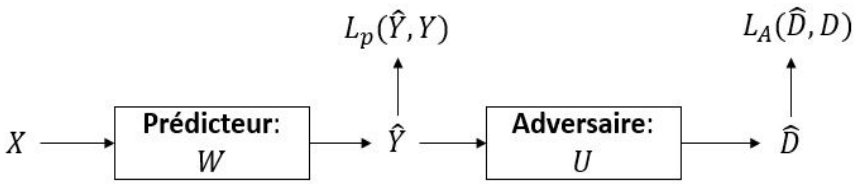

FIGURE 3.2 – Structure d'un mitigateur adverse

Le prédicteur reçoit les données d'entrée X et produit une prédiction $\hat{Y}$, cherchant à minimiser une fonction de coût $\mathcal{L}_P(Y, \hat{Y})$ qui évalue l'écart entre les prédictions et les véritables valeurs de Y. En parallèle, l'adversaire utilise les prédictions $\hat{Y}$ comme entrée et tente de prédire la variable protégée D. Selon la définition d'équité visée, l'adversaire peut également utiliser des informations supplémentaires comme la vraie valeur de Y. Par exemple, pour atteindre l'égalité des chances, l'adversaire prend en compte à la fois $\hat{Y}$ et Y.

Le problème global d'optimisation se formalise, dans le cas de la parité démographique, comme un problème en selle :

$$\min_W \max_A \mathcal{L}_P(\hat{Y}, Y) - \alpha \mathcal{L}_A(\hat{D}, D) = \min_W \max_A \mathcal{L}_P(W(X), Y) - \alpha \mathcal{L}_A(U(W(X)), D),$$

où $\mathcal{L}_P$ représente la fonction de coût du prédicteur, $\mathcal{L}_A$ celle de l'adversaire, et α est un hyperparamètre contrôlant l'importance de la pénalisation imposée par l'adversaire. Cette formulation garantit que le prédicteur apprend à minimiser son erreur tout en réduisant la dépendance non souhaitée entre $\hat{Y}$ et D, grâce à l'opposition exercée par l'adversaire.
L'entraînement de ce modèle est réalisé de manière itérative en utilisant la descente de gradient stochastique. Dans un premier temps, l'adversaire est mis à jour pour améliorer sa capacité à prédire D en minimisant $\mathcal{L}_A$. Ensuite, le prédicteur est mis à jour pour minimiser $\mathcal{L}_P$, mais avec une pénalisation supplémentaire basée sur les gradients de l'adversaire. Concrètement, les gradients utilisés pour mettre à jour les paramètres du prédicteur sont ajustés comme suit :

$$\nabla_W \mathcal{L}_P - \text{proj}_{\nabla_W \mathcal{L}_A} \nabla_W \mathcal{L}_P - \alpha \nabla_U \mathcal{L}_A,$$

Le premier terme $\nabla_W \mathcal{L}_P$ oriente les poids du prédicteur dans la direction qui réduit l'erreur de prédiction sur la variable cible Y. Sans les termes liés à l'adversaire, c'est la direction classique utilisée dans les algorithmes de descente de gradient. Le deuxième terme quant à lui est dédié à empêcher le prédicteur de se déplacer dans une direction qui pourrait involontairement *aider l'adversaire* à réduire sa propre fonction de coût. Le dernier terme introduit une pénalisation explicite pour réduire la performance de l'adversaire. En multipliant $\nabla_{\theta_f} \mathcal{L}_A$ par un hyperparamètre réglable $\alpha > 0$, le prédicteur est poussé à ajuster ses poids dans une direction qui augmente la perte de l'adversaire, rendant D plus difficile à prédire à partir de $\hat{Y}$

Avantages et limites de la méthode Cette méthode présente plusieurs avantages. Elle est tout d'abord très flexible, car elle peut être appliquée à diverses définitions d'équité comme la parité démographique ou l'égalité des chances. De plus, elle est agnostique au modèle du prédicteur, ce qui signifie qu'elle peut être utilisée avec tout modèle entraîné par des méthodes basées sur les gradients. Enfin, elle garantit théoriquement que, sous certaines conditions, le modèle obtenu satisfait les contraintes d'équité tout en maintenant une performance acceptable sur la tâche principale.

La méthode comporte cependant des limitations importantes. L'entraînement simultané des deux réseaux peut être instable et nécessite un réglage précis des hyperparamètres pour éviter des problèmes de divergence. La présence de deux réseaux accroît également la complexité

computationnelle, rendant l'entraînement plus coûteux en temps et en ressources. Enfin, dans le domaine de l'assurance, l'utilisation de réseaux neuronaux peut poser des problèmes en raison de leur faible interprétabilité, ce qui limite leur adoption dans des environnements où la transparence est essentielle.

3.3 Méthodes après la modélisation (post-traitement)

Les techniques post-traitement consistent à ajuster les prédictions d'un modèle déjà entraîné afin de corriger les biais ou d'améliorer l'équité. Ces méthodes sont particulièrement utiles lorsque l'algorithme sous-jacent ne peut pas être modifié ou que sa modification serait très complexe à mettre en oeuvre.

3.3.1 Mitigation par les barycentres de Wasserstein

Les barycentres de Wasserstein représentent une méthode pour ajuster les distributions de prédictions de manière à garantir une équité statistique, tout en respectant la structure sous-jacente des données. Contrairement aux approches classiques de rééchelonnement, cette méthode prend en compte non seulement les différences moyennes entre les groupes, mais aussi l'ensemble de la distribution des prédictions. Elle est proposée par A. Charpentier et al. (2023) [6].

Définition et formulation mathématique

Soient deux groupes D_1 et D_2 définis par une variable sensible D, et $f(x, D = D_2)$ et $f(x, D = D_1)$ les prédictions correspondantes pour chaque groupe. L'objectif est de construire une nouvelle distribution de prédictions, appelée *barycentre de Wasserstein*, qui combine les deux distributions tout en satisfaisant un critère d'équité. Mathématiquement, le barycentre est défini comme la solution du problème d'optimisation suivant :

$$P^\star = argmin_Q \sum_{i \in \{D_1, D_2\}} \omega_i W_2^2(Q, P_i),$$

où :

— P_i représente la distribution des prédictions du groupe i,

— $W_2^2(Q, P_i)$ est la distance de Wasserstein au carré entre Q et P_i,

— $\omega_i \geq 0$ sont des poids assignés à chaque groupe, tels que $\sum_i \omega_i = 1$.

En dimension unidimensionnelle, la solution au problème peut être simplifiée grâce aux fonctions de répartition F_i et à leurs inverses F_i^{-1} ([6]). Le barycentre est alors donné par :

$$F^\star(x) = \Big(\sum_{i \in \{A,B\}} \omega_i F_i^{-1} \circ F_1(x) \Big) F_1,$$

et les nouvelles prédictions sont obtenues par une moyenne pondérée des anciennes et d'une transformation monotone des anciennes :

$$m^\star(x, s = D_i) = \mathbb{P}(D = D_i).m(x, D = D_i) + \mathbb{P}(D = D_j).F_j \circ F_i^{-1}(m(x, s = D_i)), i \neq j.$$

Application aux modèles de tarification

Cette méthode est appliquée pour réduire les disparités entre les prédictions des groupes D_1 (par exemple, hommes) et D_2 (par exemple, femmes). Prenons un exemple de score prédictif dans l'assurance automobile. Si la probabilité prédite pour D_1 est plus élevée que pour D_2, un simple rééchelonnement pourrait réduire cet écart moyen, mais ne traiterait pas les différences plus complexes dans les queues de la distribution. Les barycentres de Wasserstein, en revanche, ajustent les prédictions tout en préservant les caractéristiques globales des distributions.

Propriétés et avantages

La méthode des barycentres de Wasserstein présente plusieurs avantages :

— **Equilibre des distributions :** Les nouvelles prédictions $m^\star(X)$ satisfont le principe d'équilibre selon lequel les prédictions égalisent la variable cible en moyenne.

— **Conservation de la structure :** Contrairement à un simple rééchelonnement, cette méthode respecte la structure statistique des distributions d'origine.

— **Flexibilité :** Elle peut être adaptée pour différents types de modèles et de distributions, et permet une interprétation intuitive grâce à la transformation monotone des scores.

Limites et considérations

Toutefois, cette méthode n'est pas sans limites. Elle repose sur une estimation précise des distributions P_{D_i} , ce qui peut être difficile en présence de données rares ou bruitées. De plus, bien que la méthode garantisse une équité statistique, elle ne répond pas nécessairement aux préoccupations éthiques plus larges liées à l'équité, comme le principe de suffisance ou la séparation.

3.3.2 Le transport optimal

Le principe est le même que celui décrit dans les approches au niveau des données 3.1. L'idée est de projeter chacune des distributions conditionnelles des prédictions dans une indépendante de D.

L'avantage ici est qu'en général, à l'étape post-modélisation la variable d'intérêt est unidimensionnelle. Dans le cas d'une prédiction continue, on a une formule fermée pour la transformation, une fois la distribution cible F^* choisie ([21]).

$$\hat{Y}_d^* = F^* \circ F_d(\hat{Y}_d),\ \hat{Y}_d^* \text{ prédiction transformée}$$

Tout comme pour la méthode précédente, la nécessité ici est d'estimer les lois conditionnelles F_d, ce qui peut être complexe si les données ne sont pas en volume important. Le problème majeur est le choix de la distribution cible. C'est pourquoi, la méthode précédente est préférée.

3.3.3 Optimisation des seuils (*Threshold Optimization*)

L'optimisation des seuils, ou *Threshold Optimization*, est une méthode particulièrement adaptée aux modèles de classification. Contrairement aux techniques de recalibrage probabiliste classiques, cette approche vise à ajuster les seuils de décision de manière à satisfaire exactement une contrainte d'équité prédéfinie, telle que la parité du taux de vrais positifs ou la parité des erreurs de classification entre différents groupes sensibles.

Principe de fonctionnement

L'idée de base est de définir des seuils spécifiques pour chaque sous-groupe de la variable sensible, afin de garantir que les métriques d'équité soient respectées. Par exemple, dans le cadre d'un modèle binaire, chaque prédiction est convertie en une décision (0 ou 1) en comparant la probabilité prédite à un seuil. Au lieu d'utiliser un seuil unique pour tous les groupes, l'optimisation des seuils détermine un seuil optimal pour chaque groupe, en respectant les contraintes d'équité spécifiées.

La méthode consiste, pour chaque valeur de la variable sensible, à :

— **Calculer les seuils**

On génère tous les seuils possibles et sélectionne la combinaison optimale en fonction d'un objectif principal (par exemple, maximiser la précision) tout en respectant les contraintes d'équité.

— **Randomiser entre deux seuils**

Pour garantir une stricte satisfaction des contraintes, il peut être nécessaire de randomiser les prédictions entre deux seuils proches, selon une probabilité calculée.

Exemple : Parité du taux de vrais positifs

Supposons que la contrainte d'équité impose la parité du taux de vrais positifs (*True Positive Rate Parity*). Dans ce cas : 1. La méthode explore les courbes de performances des groupes sensibles en fonction des seuils. 2. Elle identifie les seuils où les courbes respectent la contrainte d'équité tout en maximisant l'objectif global, comme l'exactitude ou une autre métrique de performance. 3. Le point optimal est choisi de manière à pondérer les contributions des groupes sensibles en fonction de leur taille respective.

Limites

Bien que l'optimisation des seuils permette de satisfaire exactement les contraintes d'équité, elle peut entraîner une baisse de la performance globale, notamment en termes de précision. De plus, cette méthode nécessite la disponibilité de la variable sensible au moment de la prédiction, ce qui peut poser des défis en termes de confidentialité ou de faisabilité opérationnelle.

En résumé, les méthodes de mitigation de l'iniquité des modèles sont variées et doivent être choisies en fonction des objectifs et des contraintes spécifiques du modèle. Les méthodes de prétraitement sont relativement simples, mais peuvent ne pas suffire à éliminer les biais subtils. Les techniques pendant la modélisation offrent plus de contrôle sur les objectifs d'équité, mais augmentent la complexité de l'apprentissage. Enfin, les méthodes de post-traitement permettent de corriger les biais après coup, mais peuvent parfois entraîner une perte de précision. Le choix de la méthode la plus adaptée dépendra donc du type de données, des objectifs de modélisation et des contraintes légales ou éthiques en vigueur.

Chapitre 4

Evaluation de l'équité des modèles : Application à l'assurance automobile

Dans les chapitres précédents, l'examen des définitions et des métriques d'équité a permis d'identifier et d'étudier différentes méthodes de mitigation des inéquités dans les modèles d'apprentissage. À ce stade, ces concepts seront appliqués dans le cadre d'un cas pratique : la tarification en assurance automobile. Cette application vise à évaluer les biais présents dans les données, à mesurer l'équité des modèles prédictifs, et à explorer les possibilités de correction des discriminations observées.

La première étape de cette étude consiste à construire un modèle de tarification basé sur l'approche classique *fréquences-coûts*. Cependant, l'accent sera mis sur une analyse approfondie des données, notamment pour détecter les déséquilibres liés à la variable protégée `DRIVER_GENDER`. Une exploration des relations multivariées entre les variables explicatives permettra de mettre en évidence des corrélations explicites ou implicites susceptibles d'affecter l'équité des modèles.

Enfin, la construction des modèles prédictifs pour la prime pure reposera sur deux approches complémentaires : le modèle linéaire généralisé (GLM), le modèle $LightGBM$ basé sur des arbres de décision. Ces modèles seront comparés en termes de performance, d'interprétabilité et de respect des critères d'équité.

Ce chapitre est organisé comme suit :

— La Section 4.1 présente la base de données et ses premières caractéristiques, en soulignant les éventuelles sources de déséquilibres ou biais.

— La Section 4.2 propose une analyse multivariée des variables, avec une attention particulière portée aux corrélations explicites ou cachées avec la variable protégée `driver_gender`.

— La Section 4.4 décrit la construction et l'évaluation des modèles prédictifs GLM, $LightGBM$ et GAM pour la prédiction de la prime pure.

L'objectif est d'appliquer une démarche méthodologique rigoureuse pour combiner précision prédictive et équité, tout en explorant les impacts des biais présents dans les données et les modèles.

4.1 Présentation de la base de données et des premières caractéristiques des assurés

4.1.1 Présentation de la base de données

La base de données utilisée dans cette étude provient d'un portefeuille d'assurance automobile « responsabilité civile » d'un assureur belge en 1997. Elle comprend 163 231 souscripteurs uniques, chacun observé sur une période allant d'un jour à une année. Les caractéristiques

des assurés, des véhicules, des contrats, ainsi que les informations spatiales sont disponibles, regroupées en plusieurs catégories.

Le tableau 4.1 détaille les variables disponibles, leurs descriptions et leurs types.

TABLE 4.1 – Description des variables disponibles dans la base de données

Variable	Description	Type
Variables d'intérêt		
nclaims	Nombre de sinistres déclarés par l'assuré	Catégorielle
exp	Fraction de l'année pendant laquelle l'assuré était exposé au risque	Continue
amount	Montant total des sinistres déclarés (en euros)	Continue
Caractéristiques des assurés		
ageph	Âge de l'assuré en années	Continue
sex	Genre de l'assuré (homme ou femme)	Catégorielle
Caractéristiques des contrats		
coverage	Type de couverture : TPL (uniquement la responsabilité civile -rc), TPL+ (rc et dommages partiels) TPL++(rc et dommages complet)	Catégorielle
bm	Niveau du bonus-malus (de 0 à 22, plus élevé indique un historique de sinistres moins favorable)	Continue
fleet	Véhicule faisant partie d'une flotte (oui ou non)	Catégorielle
Caractéristiques des véhicules		
fuel	Type de carburant du véhicule (essence ou diesel)	Catégorielle
power	Puissance du véhicule en kilowatts	Continue
agec	Âge du véhicule en années	Continue
use	Utilisation principale du véhicule (privée ou professionnelle)	Catégorielle
Données spatiales		
long	Coordonnée de longitude du centre de la municipalité de résidence	Continue
lat	Coordonnée de latitude du centre de la municipalité de résidence	Continue
postcode	Code postal de la municipalité de résidence	catégorielle

Les données fournissent une base solide pour modéliser les sinistres en fonction des caractéristiques des assurés, des véhicules et des contrats. Les informations spatiales enrichissent cette analyse en introduisant des dimensions géographiques.

Cette base de données provient du packages R de C. Dutang et A. Charpentier (2024) [12] et a été utilisée dans d'autres études. De ce fait, elle ne présente pas d'anomalie particulière (données manquantes, typologie de variables, ...) et ne nécessite pas de traitements préliminaires importants. Notons cependant la création de deux variables nécessaires à la modélisation de la prime pure : le coût moyen (`COST`) et la fréquence (FREQ) annualisée des sinistres. Leur produit permet de construire une quantité que nous appelons 'prime pure' (`PURE_PREMIUM`), en référence aux modèles *fréquences* $\times$ *coûts*.

4.1.2 Informations générales sur le portefeuille des assurés

Cette section décrit les principales caractéristiques du portefeuille d'assurance automobile analysé, en mettant en avant les données relatives aux sinistres ainsi qu'aux facteurs de risque influençant les primes.

La base de données inclut des informations détaillées sur les sinistres déclarés par les assurés, dont la majorité (77,35%) est exposée aux risques durant une année entière. Sur l'ensemble du portefeuille, environ 88,80% des assurés n'ont déclaré aucun sinistre pendant leur période de couverture, tandis que 10,14% ont déclaré un sinistre. Seule une proportion marginale de 1,06% a signalé deux sinistres ou plus. En termes de coûts, la majorité des sinistres concerne de faibles montants ; seuls 2% des assurés ayant déclaré un sinistre ont eu un coût supérieur à 10 000 euros. La fréquence globale des sinistres est de 13,93%, calculée comme le ratio entre le nombre total de sinistres et l'exposition globale (en années). Quant à la sévérité moyenne des sinistres - définie comme le ratio entre leur coût et leur nombre totaux - elle est estimée à 1309,17 euros, .

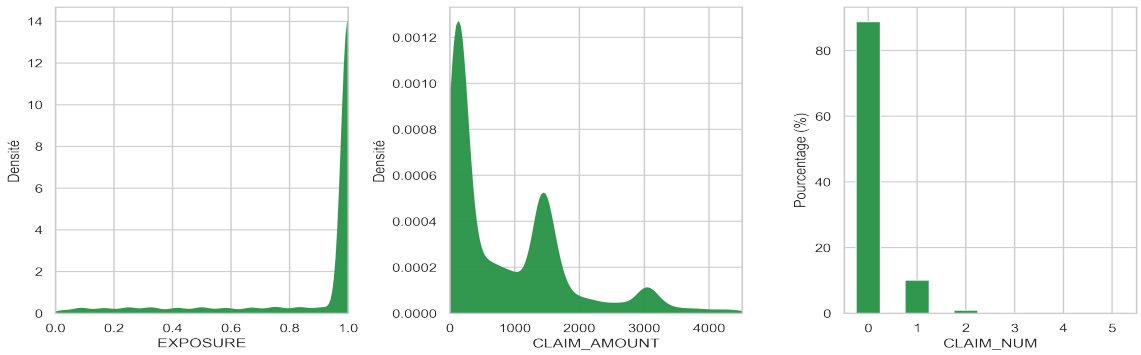

FIGURE 4.1 – Distributions de l'exposition et des sinistres

Les facteurs de risque jouent un rôle central dans la tarification. Parmi les variables catégorielles, le type de couverture choisi par l'assuré (`coverage`) est un élément crucial. La majorité des assurés (58,28%) optent pour une couverture minimale (*TPL*), tandis que les omnium partielle (garanties incendie, vol, bris de glace, heurt animal et forces de la nature [1], notation *TPL+*) et complète (garantie de tous les dégâts suite à un accident avec un autre automobiliste ou du fait de sa conduite - en droit ou en tort - et au vandalisme, notation *TPL++*, en plus des garanties de l'omnium partielle) sont respectivement choisies par 28,17% et 13,54%. Le type de carburant (`fuel`) reflète une prédominance des véhicules essence (69,12%) par rapport au diesel (30,88%). Le genre des assurés (`sex`) est également documenté, avec une majorité d'hommes (73,55%), tandis que l'utilisation des véhicules est en grande partie privée (95,17%). Une faible proportion des véhicules (3,17%) appartient à des flottes professionnelles.

Les variables continues offrent une granularité supplémentaire pour caractériser les assurés et leurs véhicules. Par exemple, l'âge des assurés (`ageph`) est largement concentré entre 25 et 75 ans (93,53%), avec des proportions plus faibles de jeunes conducteurs et de seniors. Concernant les véhicules, la puissance (`power`) se situe principalement en dessous de 100 kilowatts (97,35%), et leur ancienneté (`agec`) est inférieure à 20 ans pour 99,53% des observations. Une autre mesure clé est le niveau de bonus-malus (`bm`), où plus de la moitié des assurés se situent dans les niveaux

1. grêle, tempête/ouragan, inondations, raz-de-marées, avalanches, chutes de pierres ou de neige, tremblement de terre, éruptions volcaniques,

0 (37,77%) et 1 (16,52%). Cela reflète leur expérience de conduite et leur historique de sinistres.

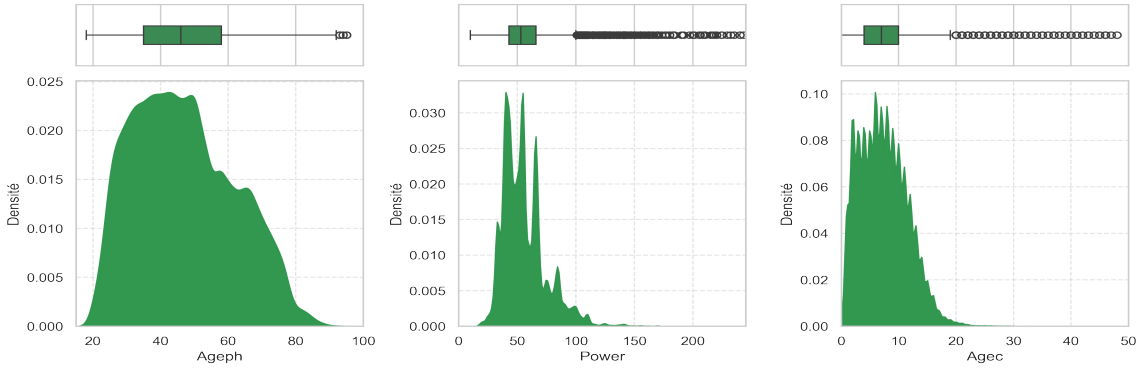

FIGURE 4.2 – Distributions des facteurs de risques continus

Ces premières descriptions mettent en lumière les caractéristiques clés du portefeuille. Cependant, une simple analyse univariée est insuffisante pour comprendre les relations sous-jacentes entre les variables et leurs implications sur les sinistres. Une analyse multivariée plus approfondie sera nécessaire pour explorer les interactions entre les facteurs de risque et identifier les corrélations, notamment celles en lien avec le genre de l'assuré (`driver_gender`). Cela permettra de mieux comprendre les sources de déséquilibres potentiels dans les données.

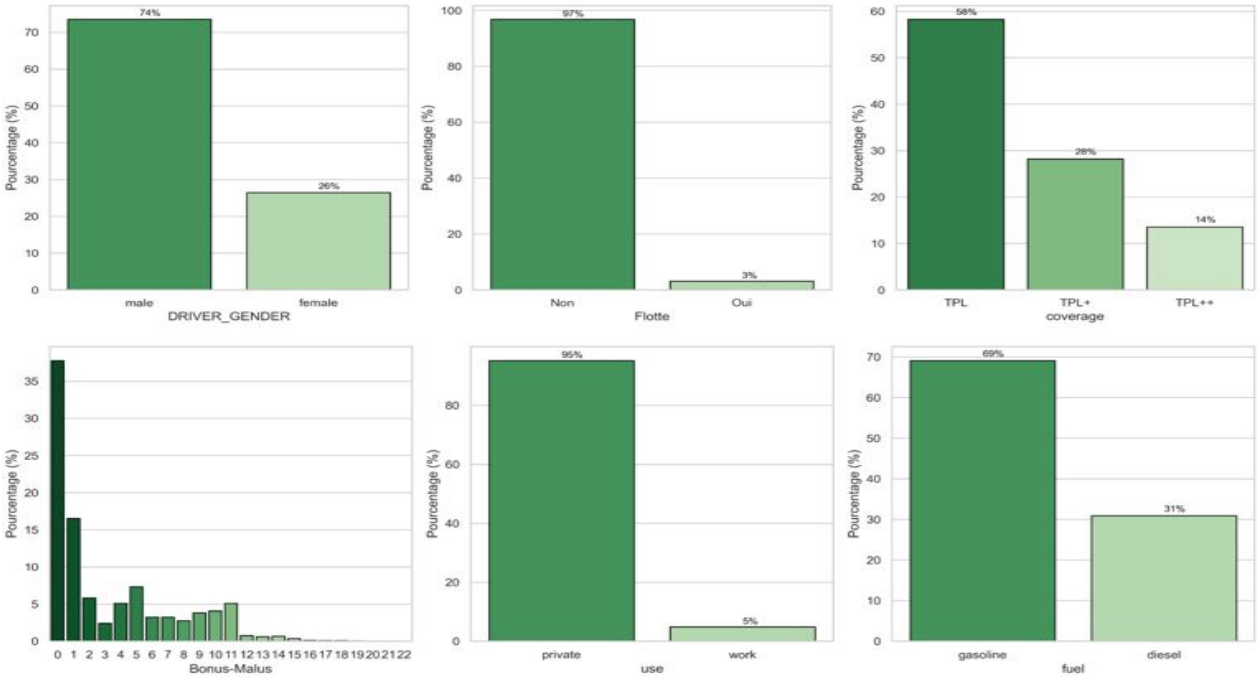

FIGURE 4.3 – Distributions des facteurs de risques catégoriels

4.2 Analyse bivariée des caractéristiques du portefeuilles

L'analyse bivariée constitue une étape essentielle pour explorer les relations entre les différentes caractéristiques du portefeuille. Elle permet de mieux comprendre les interactions potentielles entre les variables, qu'elles soient continues ou catégorielles, et d'identifier des dépendances pouvant avoir un impact sur la modélisation des primes. Dans ce cadre, des métriques adaptées ont été utilisées pour quantifier ces relations : les corrélations de Pearson, Kendall et

partielles pour les variables continues, et le V de Cramér ainsi que l'information mutuelle normalisée pour les variables catégorielles. Pour analyser les interactions entre variables continues et catégorielles, l'ANOVA (Analyse de variance) et le test de Kuskal-Wallis sont employés.

Corrélations des variables continues

L'analyse des corrélations entre les variables continues met en lumière certaines relations intéressantes et aide à comprendre les interactions potentielles entre les facteurs de risque. Les corrélations ont été évaluées en utilisant les coefficients de Pearson, Kendall et les corrélations partielles.

Parmi les variables continues, `Bonus-Malus` présente des corrélations positives faibles avec `PURE_PREMIUM` dans les matrices de Pearson ($r = 0.017$) et Kendall ($r_k = 0.062$), reflétant une légère augmentation des primes pour les conducteurs ayant un historique de sinistres moins favorable. Cependant, dans la matrice des corrélations partielles, ce lien devient négatif ($r_p = -0.013$), suggérant que cette relation est en partie influencée par d'autres facteurs, tels que les caractéristiques des véhicules ou des contrats. Une observation similaire peut être faite pour `ageph` (âge de l'assuré), où une corrélation initiale légèrement négative avec `PURE_PREMIUM` ($r = -0.015$) s'atténue presque complètement dans les corrélations partielles ($r_p = -0.004$).

Les variables `power` (puissance du véhicule) et `agec` (ancienneté du véhicule) montrent des relations très faibles avec `PURE_PREMIUM` dans toutes les matrices de corrélation, avec des coefficients proches de zéro. Cela indique que ces caractéristiques semblent n'avoir qu'une influence marginale directe sur la prime pure. Cependant, elles pourraient interagir indirectement avec des variables catégorielles telles que `coverage` ou `fuel`, qui influencent fortement la tarification.

En ce qui concerne les relations entre les facteurs de risque eux-mêmes, `ageph` (âge de l'assuré) et `Bonus-Malus` présentent une forte corrélation négative dans toutes les matrices ($r = 0.39$ pour Pearson). Cette relation peut s'expliquer par le fait que les assurés les plus âgés, bénéficiant généralement d'une plus grande expérience de conduite, ont tendance à adopter des comportements plus prudents et à cumuler des antécédents de conduite favorables. Cela se traduit par un meilleur score de bonus-malus, indiquant un risque perçu plus faible par l'assureur. À l'inverse, les jeunes conducteurs, souvent moins expérimentés, peuvent présenter un historique de sinistres plus élevé, reflété par un niveau de bonus-malus moins avantageux. De même, `agec` et `power` montrent une faible corrélation négative ($r = -0.212$), indiquant que les véhicules plus récents ont tendance à être plus puissants.

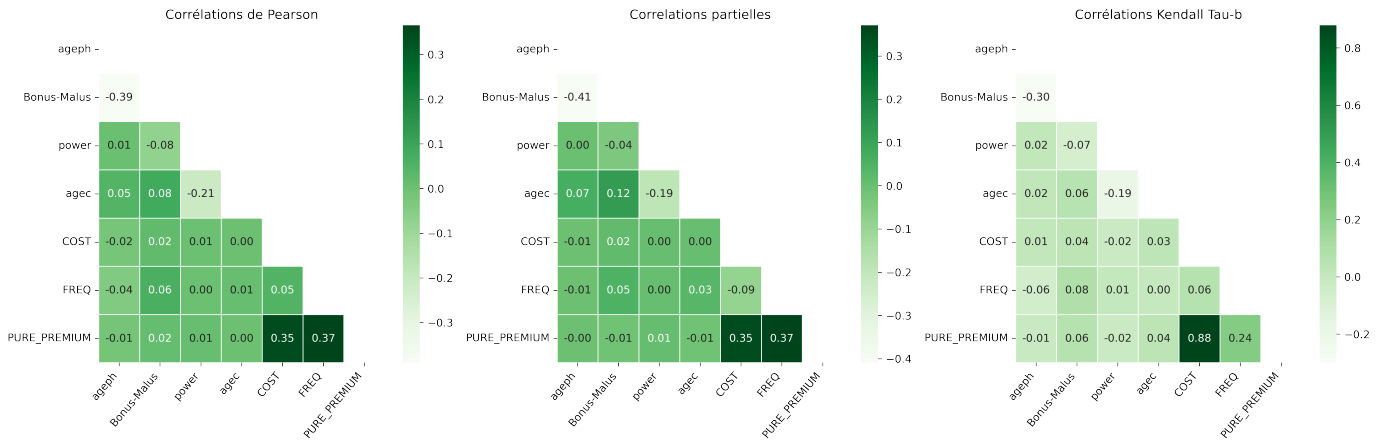

FIGURE 4.4 – Dépendance entre facteurs de risque continus

Dépendance entre facteurs de risque catégoriels

L'évaluation de la dépendance entre les variables catégorielles repose sur deux métriques complémentaires : le V de Cramér, qui mesure la force des associations entre variables qualitatives en s'appuyant sur le Khi-deux, et l'information mutuelle normalisée, qui quantifie le partage d'information entre deux variables, captant potentiellement des relations plus complexes ou non linéaires.

Test du Khi-deux et V de Cramer

Le **test du Khi-deux d'indépendance** [14] est utilisé pour évaluer s'il existe une association significative entre deux variables catégorielles. Il teste l'hypothèse nulle (H_0) selon laquelle les deux variables sont indépendantes, c'est-à-dire qu'il n'y a pas de relation entre elles.
Pour réaliser ce test, on construit un tableau de contingence qui répertorie les fréquences observées pour chaque combinaison des modalités des deux variables. Ensuite, on calcule les fréquences attendues sous l'hypothèse d'indépendance, à l'aide de la formule :

$$E_{ij} = \frac{N_i \cdot N_j}{N},$$

où E_{ij} est la fréquence attendue pour la cellule (i, j), N_i et N_j sont les totaux marginaux des lignes et des colonnes, et N est le nombre total d'observations.
La statistique du Khi-deux est ensuite calculée comme suit :

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

où O_{ij} est la fréquence observée dans la cellule (i, j). Cette statistique suit une loi du Khi-deux avec un nombre de degrés de liberté égal à $(r-1)(c-1)$, où r est le nombre de lignes et c le nombre de colonnes du tableau de contingence. Si la statistique du Khi-deux est suffisamment grande, on rejette l'hypothèse d'indépendance.

Le **V de Cramer**[15] quant à lui indique l'intensité de la dépendance entre deux variables catégorielles. Il s'obtient en normalisant convenablement la statistique du test de Khi-deux. Formellement il s'écrit :

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, r-1)}}$$

où χ^2 est la statistique du test de Khi-deux, n est la taille de l'échantillon, k est le nombre de colonnes, et r est le nombre de lignes dans le tableau de contingence. Le **V de Cramer** prend des valeurs entre 0 et 1, où 0 indique une indépendance totale entre les deux variables, et 1 indique une relation de dépendance parfaite. Cet indicateur est particulièrement utile pour comparer l'intensité de la relation entre deux variables, quel que soit le nombre de catégories dans celles-ci.

Les coefficients V de Cramér révèlent globalement de faibles niveaux de dépendance entre les variables catégorielles. Par exemple, `coverage` présente une dépendance modérée avec `use` ($V = 0.1016$) et `fuel` ($V = 0.0948$), ce qui pourrait indiquer que certains types de couverture sont privilégiés en fonction de l'utilisation ou du type de carburant des véhicules. De manière similaire, `DRIVER_GENDER` affiche une corrélation faible avec `fuel` ($V = 0.1034$), reflétant peut-être des préférences marginales pour certains types de carburant selon le genre des assurés. En revanche, la variable `Flotte` montre des dépendances négligeables avec toutes les autres variables ($V < 0.094$), confirmant son rôle relativement indépendant dans le jeu de données.

L'information mutuelle normalisée renforce ces observations, tout en indiquant une intensité

plus faible des relations. Par exemple, l'association entre `coverage` et `use`, qui était notable dans le V de Cramér ($V = 0.1016$), est plus atténuée avec l'information mutuelle ($I = 0.0075$). Cette différence pourrait s'expliquer par une structure moins linéaire de la relation. De même, la relation entre `DRIVER_GENDER` et `fuel`, identifiée comme légèrement significative dans le V de Cramér ($V = 0.1034$), est également détectée ici ($I = 0.0093$). Enfin, `Flotte`, qui affiche des niveaux très faibles de dépendance dans le V de Cramér, conserve des coefficients quasi nuls dans l'information mutuelle ($I < 0.017$).

En combinant les deux métriques, il est possible de mieux comprendre la nature des relations entre les variables. Tandis que le V de Cramér capture les relations globales en se basant sur la fréquence conjointe des catégories, l'information mutuelle permet de détecter des relations non linéaires ou subtiles. Ensemble, elles confirment que les dépendances entre les variables catégorielles de cette base de données sont généralement faibles, à l'exception de quelques interactions modérées impliquant `coverage`, `use` et `fuel`.

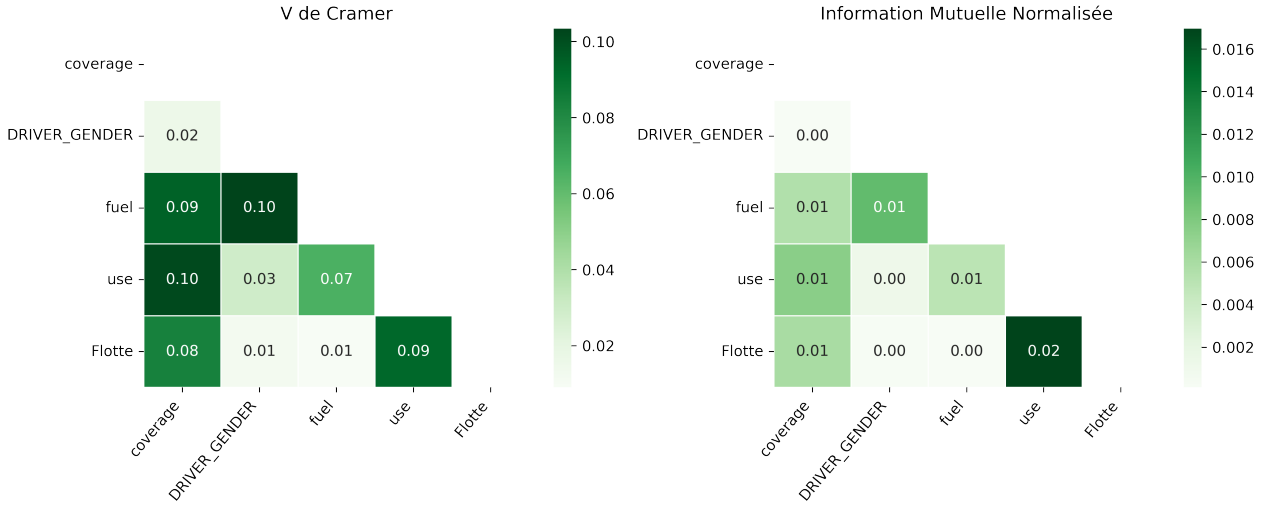

FIGURE 4.5 – Dépendance entre facteurs de risque catégoriels

Dépendance entre facteurs de risques continus et catégoriels

Pour évaluer la dépendance entre les facteurs de risque continus et catégoriels, nous avons utilisé deux approches complémentaires : l'analyse de variance (ANOVA) et le test non paramétrique de Kruskal-Wallis. Ces deux outils permettent d'examiner si les distributions des variables continues diffèrent significativement entre les catégories des variables qualitatives. Les résultats des statistiques ANOVA et Kruskal-Wallis sont présentés dans les tableaux joints en annexe et fournissent des indications précieuses sur les relations sous-jacentes.

Les deux tests montrent que certaines relations entre variables continues et catégorielles sont particulièrement marquées. Par exemple, `agec` (ancienneté du véhicule) affiche des dépendances très fortes avec `coverage` et `fuel`, avec des statistiques élevées dans les deux cas (respectivement 20989.64 et 37749.60 pour ANOVA, 2651.19 et 2651.20 pour Kruskal-Wallis). Cela reflète une tendance selon laquelle les véhicules les plus neufs et les plus puissants sont associés à des couvertures d'assurance beacoup plus complètes.

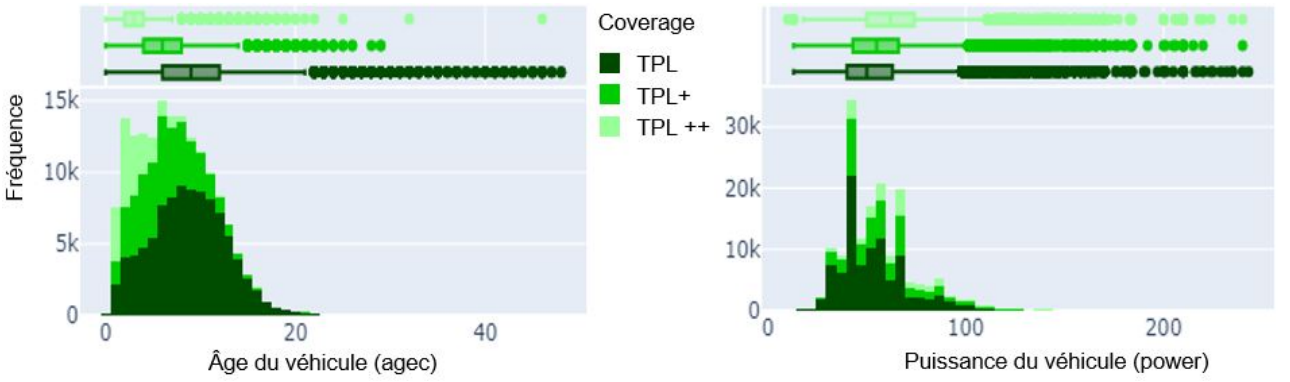

FIGURE 4.6 – Distributions de l'âge et de la puissance des véhicules conditionnellement au type de couverture

De manière similaire, `power` (puissance du véhicule) montre une forte association avec `use` (usage du véhicule) et `DRIVER_GENDER`, indiquant que les véhicules professionnels ou appartenant à certains profils d'assurés ont souvent des puissances plus élevées.
En revanche, certaines relations sont beaucoup moins significatives. Par exemple, les variables `COST`, `FREQ`, et `PURE_PREMIUM` montrent peu de dépendances avec des facteurs tels que `Flotte` et `DRIVER_GENDER`, comme l'illustrent des statistiques ANOVA et Kruskal-Wallis proches de zéro (< 1.0 dans plusieurs cas). Cela suggère que ces caractéristiques catégorielles n'ont pas d'influence directe sur les coûts ou fréquences des sinistres.
Des différences mineures sont observées entre les résultats des deux méthodes. Le test de Kruskal-Wallis, qui est moins sensible aux écarts à la normalité, identifie certaines relations légèrement plus significatives, comme entre `ageph` (âge de l'assuré) et `use` (type d'usage de véhicules, où les statistiques augmentent de 77.81 (ANOVA) à 116.95 (Kruskal-Wallis). Cela pourrait indiquer des effets non linéaires ou asymétriques dans la relation entre ces variables. Mais dans l'ensemble, les deux tests sont en cohérence.

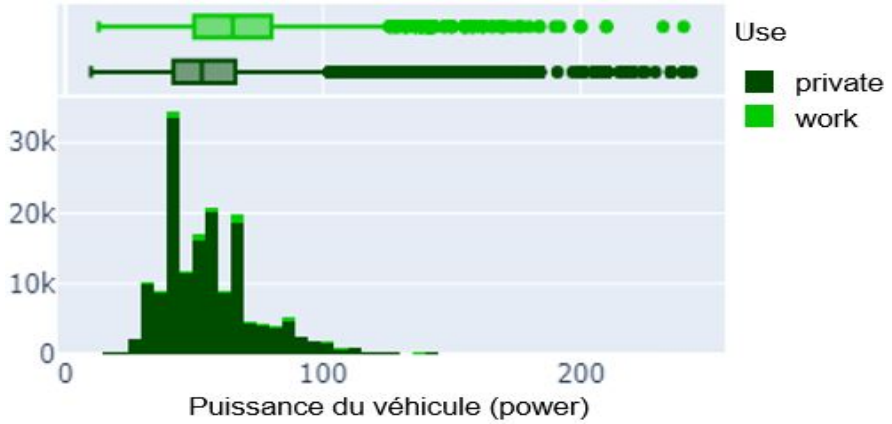

FIGURE 4.7 – Distribution de la puissance conditionnellement au type d'usage

Les analyses descriptives précédentes (uni et bidimensionnelles) permettent de ressortir les informations suivantes sur le portefeuille des assurés :

— Les analyses unidimensionnelles montrent que la majorité des assurés n'ont déclaré aucun sinistre (88,80%) et que seuls 2% des sinistres impliquent des coûts supérieurs à 10 000 euros. Les variables continues révèlent une concentration notable : 93,53% des assurés ont un âge compris entre 25 et 75 ans, et 97,35% des véhicules ont une puissance inférieure à 100 kilowatts.

— Concernant les variables catégorielles, la majorité des assurés (58,28%) optent pour une couverture minimale (*TPL*), et l'utilisation des véhicules est principalement privée (95,17%). Le genre des assurés (`DRIVER_GENDER`) est déséquilibré, avec une majorité d'hommes (73,55%).

— Les analyses bidimensionnelles montrent des corrélations modérées entre certaines variables continues. Par exemple, `ageph` (âge de l'assuré) et `Bonus-Malus` présentent une corrélation négative ($r = -0.39$), reflétant l'effet d'expérience des conducteurs. En revanche, les relations entre `PURE_PREMIUM` et des variables comme `power` ou `agec` sont très faibles, suggérant une influence indirecte limitée.

— En étudiant les relations entre les variables catégorielles et continues, des associations significatives sont détectées. Par exemple, `agec` (ancienneté du véhicule) montre une forte dépendance avec `coverage`, indiquant une préférence pour des couvertures plus complètes pour les véhicules plus récents. Toutefois, `DRIVER_GENDER` présente une faible dépendance avec la majorité des variables catégoriques et continues.

Ces observations soulignent des caractéristiques marquantes du portefeuille, mais posent également la question de l'équité. Existe-t-il des disparités historiques liées au genre (`DRIVER_GENDER`) dans les données ? Les analyses suivantes s'intéresseront à cette problématique en explorant la parité dans les sinistres et l'impact potentiel des proxys.

Analyse de variance (anova) et test de Kruskal-Wallis

L'analyse de variance (analysis of variances en anglais, ou anova) et le test de **Kruskal-Wallis** sont des outils statistiques cruciaux pour comparer les distributions de variables continues entre plusieurs groupes définis par une variable qualitative. Ces deux approches visent à détecter des différences significatives dans les moyennes ou les distributions, tout en reposant sur des hypothèses différentes.

Analyse de variance (anova)

L'**anova**[31] repose sur deux hypothèses fondamentales : la normalité des données dans chaque groupe et l'homogénéité des variances entre les groupes (homoscédasticité). Elle compare la variance expliquée par les différences entre les groupes (variance intergroupes) à la variance inexpliquée à l'intérieur des groupes (variance intragroupes).
La variance **intragroupes** mesure la dispersion des observations au sein de chaque groupe :

$$\mathrm{SS}_{\text{intra}} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2,$$

où g est le nombre de groupes, n_i est le nombre d'observations dans le groupe i, x_{ij} est la valeur de l'observation j dans le groupe i, N le nombre total d'observations et $\bar{x}_i$ est la moyenne du groupe i.
La variance **intergroupes** mesure la dispersion entre les moyennes des groupes :

$$\mathrm{SS}_{\text{inter}} = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})^2,$$

où $\bar{x}$ est la moyenne générale (tous groupes confondus).
La statistique F de l'ANOVA est donnée par :

$$F = \frac{\mathrm{MS}_{\text{inter}}}{\mathrm{MS}_{\text{intra}}},$$

où $\mathrm{MS}_{\text{inter}} = \frac{\mathrm{SS}_{\text{inter}}}{g-1}$ et $\mathrm{MS}_{\text{intra}} = \frac{\mathrm{SS}_{\text{intra}}}{N-g}$.

Un F élevé suggère que les différences entre les moyennes des groupes ne sont probablement pas dues au hasard. Si la statistique F dépasse une certaine valeur critique déterminée par la loi de Fisher avec $(g-1, N-g)$ degrés de liberté, l'hypothèse nulle (H_0 : égalité des moyennes) est rejetée.

Test de Kruskal-Wallis

Le **test de Kruskal-Wallis[17]** est une alternative non paramétrique à l'ANOVA. Il est adapté lorsque les hypothèses de normalité et d'homoscédasticité ne sont pas respectées. Ce test est basé sur les rangs des observations plutôt que sur leurs valeurs brutes.
La statistique du test est donnée par :

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i \left(R_i - \frac{N+1}{2} \right)^2,$$

où R_i est la somme des rangs des observations dans le groupe i, n_i est la taille du groupe i, et N est la taille totale de l'échantillon.
Sous l'hypothèse nulle (H_0 : les rangs moyens des groupes sont égaux), H suit approximativement une loi du Khi-deux avec $g-1$ degrés de liberté. Ce test est particulièrement utile lorsque les données sont ordinales ou ne respectent pas les conditions nécessaires à l'ANOVA.

Avantages et limites

- **anova** : Plus puissante lorsque les hypothèses sont respectées, mais sensible aux violations de la normalité et de l'homogénéité des variances.
- **Kruskal-Wallis** : Robuste face aux écarts à la normalité et aux hétérogénéités des variances, mais moins puissant pour des échantillons très petits ou lorsque les tailles de groupes sont déséquilibrées.

Ces deux outils se complètent pour analyser les relations entre variables continues et catégoriques, en offrant une approche statistique solide pour évaluer les proxys potentiels dans les données.

4.3 Disparités par rapport au genre du conducteur

Cette section explore les inégalités potentielles dans les données en fonction de la variable sensible `DRIVER_GENDER`. Elle vise à déterminer si des disparités historiques existent entre les hommes et les femmes en termes de fréquence et de gravité des sinistres. En parallèle, une attention particulière est portée à l'impact des variables dites « légitimes », afin d'identifier si certaines d'entre elles agissent comme des proxys pour le genre. Une telle analyse est essentielle pour anticiper et corriger d'éventuelles sources d'inéquité dans les modèles prédictifs.

4.3.1 Parité dans la fréquence et la sévérité des sinistres

Pour analyser l'équité des données, une première exploration visuelle est réalisée. Les distributions conditionnelles de la fréquence et de la sévérité des sinistres en fonction du genre du conducteur révèlent des différences légères. Comme le montrent les graphes de transport (voir figure 4.8), les distributions pour les hommes et les femmes sont très similaires sur la majeure partie de leur étendue. Les divergences apparaissent principalement au niveau des queues des distributions. Ces résultats suggèrent que, pour la majorité des sinistres, hommes et femmes présentent des comportements comparables en termes de fréquence et de coût des incidents.

Cependant, des différences émergent dans les cas extrêmes : les hommes ont tendance à générer des sinistres plus coûteux que ceux des femmes, tandis-que les fréquences sont plus importantes chez les femmes.

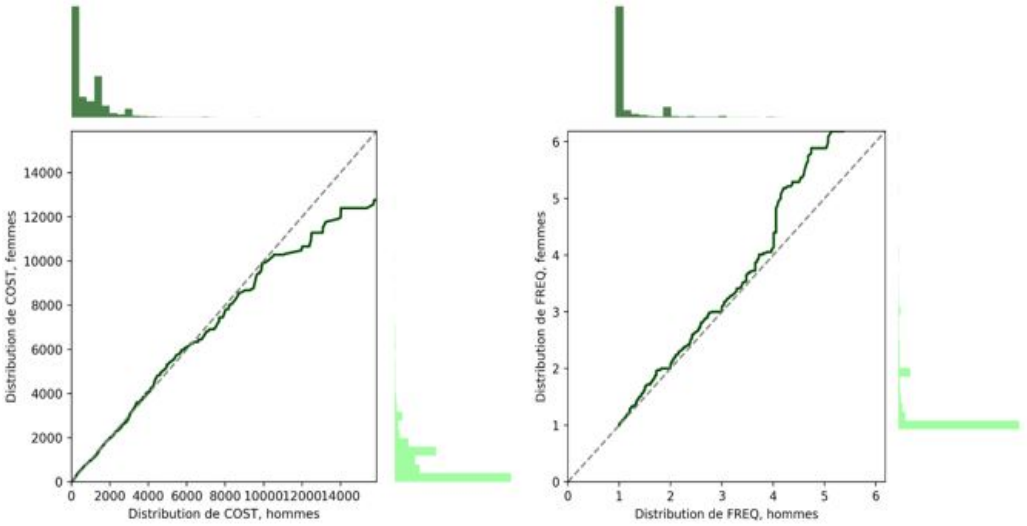

FIGURE 4.8 – Correspondances des distributions conditionnelles de coûts et de fréquences

Un examen des métriques relatives à la parité démographique, notamment la différence de parité (SPD), la distance de Wasserstein, la distance de Kolmogorov, et la métrique UF (Unfairness Metric), permet de confirmer les observations issues des distributions conditionnelles. Ces indicateurs quantifient les disparités entre les genres pour trois variables clés : la fréquence annualisée des sinistres, la gravité moyenne des sinistres, et la prime pure. En complément, les p-valeurs des tests statistiques de Kolmogorov-Smirnov et de Welch [2] évaluent la significativité des différences observées. Les résultats obtenus sont présentés dans le tableau ci-dessous et fournissent une base pour analyser en détail le degré d'équité des données.

Les résultats indiquent que pour les trois variables étudiées, les disparités entre genres sont minimes. Par exemple, la différence de parité (SPD) pour la gravité des sinistres est de $3,77\%$, un écart relativement faible. La distance de Wasserstein, qui mesure la différence globale entre les distributions, est également modérée pour cette variable (68,26). Cependant, la p-valeur du test de Kolmogorov-Smirnov ($p = 0.049$) est juste en dessous du seuil de 0,05, suggérant une légère différence statistiquement significative dans la gravité des sinistres. Pour la fréquence des sinistres et la prime pure, toutes les métriques montrent des écarts encore plus faibles, et les p-valeurs associées ($p > 0.05$) ne permettent pas de rejeter l'hypothèse d'équité.

TABLE 4.2 – Métriques et p-valeurs pour l'évaluation de la parité démographique

Métrique	**Fréquence**	**Coûts**	**Prime pure**
SPD	0.0120	0.0377	0.0254
Kolmogorov (distance)	0.0187	0.0224	0.0195
UF_metric	0.0001	0.0000	0.0000
Wasserstein	0.0143	68.26	61.61
p-valeur (Kolmogorov)	0.1515	0.0492	0.1200
p-valeur (Welch test)	0.0663	0.3130	0.6117

En résumé, les métriques confirment une quasi-absence de disparités significatives dans les données, et même dans les 'primes pures' historiques, bien que quelques différences marginales

2. Test de comparaison des moyennes en absence d'homoscédasticité

soient détectées pour la gravité des sinistres. Ces résultats suggèrent que les données présentent un bon niveau d'équité initial, ce qui constitue une base favorable pour la construction et l'évaluation des modèles prédictifs.

Cependant, cette absence apparente de disparités sur les variables cibles garantit-elle qu'il en sera de même pour les modèles ? Certaines variables légitimes pourraient-elles agir comme des proxys du genre, introduisant ainsi un biais indirect dans les prédictions ? Ces questions sont explorées dans la section suivante.

4.3.2 Les variables légitimes sont-elles des proxys du genre ?

Cette sous-section s'intéresse à la possibilité que certaines variables non sensibles contiennent implicitement de l'information sur le genre `DRIVER_GENDER`. Une telle situation pourrait entraîner un biais indirect dans les modèles prédictifs, compromettant l'équité. Pour cela, nous analysons les dépendances entre le genre du du conducteur et les autres variables explicatives, en distinguant les variables catégorielles et continues.

Dépendances avec les variables catégorielles

Les relations entre `DRIVER_GENDER` et les variables catégorielles ont été examinées à l'aide de trois outils : les p-valeurs du test de Khi-deux, le V de Cramér, et l'information mutuelle normalisée. Ces métriques permettent d'identifier si certaines variables légitimes sont fortement associées au genre du conducteur, révélant ainsi des proxys potentiels.

Les p-valeurs du test de Khi-deux montrent des associations statistiquement significatives entre `DRIVER_GENDER` et toutes les variables catégorielles analysées ($p < 0.01$). Parmi elles, la variable `fuel` (type de carburant) présente la plus forte dépendance ($p = 0.00$), suivie de `use` (usage du véhicule, $p = 6.98 \times 10^{-33}$), tandis que la variable `Flotte` affiche une dépendance plus faible ($p = 2.18 \times 10^{-6}$). Ces résultats indiquent que des liens existent entre le genre et ces variables, bien qu'ils ne reflètent pas nécessairement une corrélation élevée.

Les métriques V de Cramér et l'information mutuelle normalisée confirment ces observations en quantifiant l'intensité des dépendances. Par exemple, `fuel` présente la valeur la plus élevée pour V de Cramér ($V = 0.103$), ce qui reflète une faible dépendance mais notable. De même, l'information mutuelle pour `fuel` ($I = 0.0093$) est la plus élevée parmi les variables, renforçant l'idée qu'elle partage une petite quantité d'information avec `DRIVER_GENDER`. En revanche, `Flotte` et `use` affichent des dépendances beaucoup plus faibles ($V = 0.011$ et $I = 0.00019$ pour `Flotte`), ce qui indique une influence négligeable.

En résumé, bien que des associations statistiquement significatives soient détectées entre `DRIVER_GENDER` et plusieurs variables catégorielles, leur intensité reste faible selon V de Cramér et l'information mutuelle. Ces résultats suggèrent que les variables catégoriques analysées ne jouent qu'un rôle limité en tant que proxys pour le genre, bien qu'une attention particulière doive être portée à des variables comme `fuel` et `use` dans les étapes suivantes de l'analyse.

Qu'en est-il maintenant de la dépendance avec les variables continues ?

Dépendances avec les variables continues

Pour évaluer les dépendances entre `DRIVER_GENDER` et les variables continues, nous avons utilisé les tests statistiques d'ANOVA et de Kruskal-Wallis. Ces tests permettent de détecter des différences significatives dans les distributions des variables continues entre les groupes hommes et femmes, mettant en lumière des proxys potentiels pour la variable sensible. Les p-valeurs des tests indiquent une dépendance significative ($p < 0.05$) pour plusieurs variables, notamment `ageph` (âge de l'assuré), `power` (puissance du véhicule), et `Bonus-Malus`, tandis que

des variables comme `COST`, `FREQ`, et `PURE_PREMIUM` ne présentent pas de relations significatives avec le genre.

Pour quantifier l'intensité de ces dépendances, nous avons calculé la part de variance expliquée par `DRIVER_GENDER` ainsi que l'information mutuelle. Les résultats montrent que ces dépendances, bien que statistiquement significatives pour certaines variables, restent très faibles en intensité. Par exemple, la variable présentant la plus forte relation avec le genre, `power`, a une part de variance expliquée de seulement 2,63%, et une information mutuelle de 0,022. Ces valeurs indiquent que `DRIVER_GENDER` n'explique qu'une infime partie de la variabilité de `power`. De même, pour `ageph`, la part de variance expliquée est de 2,04% et l'information mutuelle est de 0,012, ce qui reflète une faible dépendance.

En revanche, des variables comme `agec` (ancienneté du véhicule) montrent des relations encore plus faibles avec le genre, avec une part de variance expliquée de seulement 0,10% et une information mutuelle de 0,0008. Ces résultats confirment que, bien que des différences statistiquement significatives soient détectées, leur intensité est négligeable, ce qui limite l'impact potentiel de ces variables comme proxys du genre.

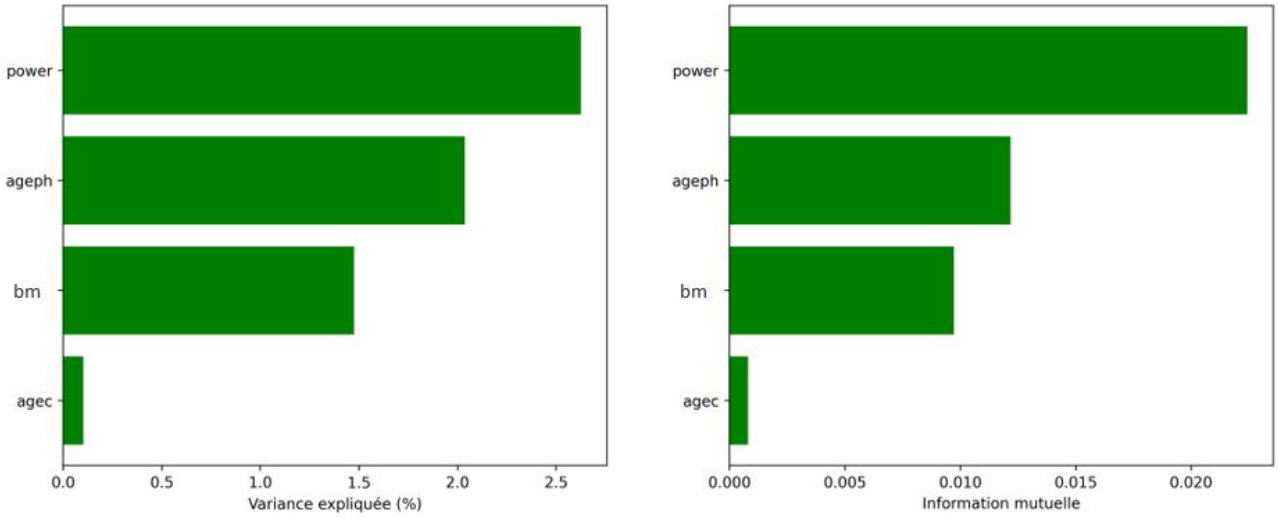

FIGURE 4.9 – Intensité de la dépendance avec le genre du conducteur

En résumé, les analyses montrent que les relations entre `DRIVER_GENDER` et les variables continues sont statistiquement significatives mais très faibles en intensité. Les variables continues analysées ne constituent donc pas de proxys majeurs pour le genre, bien qu'une attention particulière puisse être portée à `power` et `ageph` dans les analyses suivantes.

4.4 Les modèles de prime pure et leurs niveaux d'équité

Pour le calcul de la prime pure , utilisons l'approche *fréquence - coûts*. Elle découle du modèle collectif utilisé lors d'une tarification individuelle. Pour chacun des assurés i d'un portefeuilles de contrats, N^i désigne son nombre aléatoire de sinistres et $(Y_j^i)_{j>=0}$ les coûts associés. Le but est de modéliser le coût total de la sinistralité de chacun des assurés i, noté :

$$S^i = \sum_{j=1}^{N^i} Y_j^i.$$

La prime pure de l'individu i est alors obtenue en prenant la valeur moyenne de ce coût total connaissant ses caractéristiques z^i. On montre alors qu'elle est donnée par :

$$\mathbb{E}(S^i \mid Z = z^i) = \mathbb{E}(N^i \mid Z = z_i)\mathbb{E}(Y_1^i \mid Z = z_i) = fréquence \times coûts.$$

Le but de cette section est de présenter les résultats obtenus dans ce cadre pour la tarification automobile en Belgique, et d'évaluer l'équité des modèles de tarification utilisés.

4.4.1 Les modèles utilisés

Pour la modélisation séparée des fréquences et des coûts, nous utilisons deux types de modèles : les modèles linéaires généralisés (Generalized Linear Models - GLM) et les modèles de gradient boosting optimisés (Light Gradient Boosting Machine - LightGBM).

Les modèles LightGBM

LightGBM repose sur le principe du gradient boosting, une méthode d'ensemble qui combine plusieurs arbres de décision pour produire un modèle robuste et précis. Contrairement à des approches comme les forêts aléatoires, où les arbres sont indépendants, LightGBM construit chaque arbre de manière séquentielle, en s'appuyant sur les résidus des prédictions des arbres précédents.

La fonction coût minimisée par LightGBM est souvent la log-vraisemblance pour les tâches de classification ou la perte quadratique pour les régressions, mais elle peut également être adaptée pour d'autres métriques personnalisées. Par exemple, pour une tâche de régression, la fonction coût peut être exprimée comme :

$$L(\hat{y}, y) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

où y_i est la valeur observée, et $\hat{y}_i$ la prédiction. Pour une tâche de classification, on utilise généralement l'entropie croisée :

$$L(\hat{y}, y) = -\sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right].$$

Les hyperparamètres principaux de LightGBM incluent :

— **le taux d'apprentissage** (η) : contrôle l'amplitude des mises à jour des prédictions à chaque itération. Un η plus faible réduit le risque de surapprentissage mais nécessite plus d'itérations.

— **Max depth** : limite la profondeur maximale des arbres, ce qui aide à prévenir le surapprentissage.

— **Min data in leaf** : spécifie le nombre minimal de données dans une feuille, influençant la granularité des prédictions.

— **Feature fraction** : proportion de variables explicatives utilisées pour construire chaque arbre, ce qui peut améliorer la généralisation.

La méthode d'estimation de LightGBM repose sur une optimisation par gradient descendant avec régularisation. Chaque arbre est construit pour minimiser l'erreur résiduelle des prédictions précédentes en ajustant les poids attribués aux observations.

Les modèles linéaires généralisés (GLM)

Les modèles linéaires généralisés sont des extensions des modèles linéaires classiques, adaptées à des distributions non gaussiennes. La relation entre la variable réponse Y et les variables explicatives X est exprimée à l'aide d'une fonction de lien g :

$$g(\mathbb{E}[Y \mid X]) = \beta_0 + X'\beta,$$

où g est une fonction monotone adaptée à la distribution de Y. Par exemple, pour une distribution Poisson, la fonction de lien est la fonction logarithmique ($g(x) = \log(x)$).

La fonction coût minimisée dans les GLM est la log-vraisemblance négative, donnée par :

$$L(\beta) = -\sum_{i=1}^{n} \ell(y_i, \hat{y}_i),$$

où $\ell(y_i, \hat{y}_i)$ dépend de la distribution choisie. Par exemple, pour une variable réponse Y suivant une distribution de Poisson, on a :

$$\ell(y_i, \hat{y}_i) = y_i \log(\hat{y}_i) - \hat{y}_i.$$

Pour améliorer la robustesse des modèles et éviter le surapprentissage, une pénalisation Ridge (régularisation L_2) peut être ajoutée à la fonction coût :

$$L(\beta) = -\sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \lambda \sum_{j=1}^{p} \beta_j^2,$$

où λ est un hyperparamètre contrôlant la force de la régularisation.

L'estimation des paramètres dans les GLM est effectuée à l'aide de la maximisation de la vraisemblance, souvent par des méthodes numériques telles que l'algorithme IRLS (Iteratively Reweighted Least Squares).

Comparaison et spécificités

Les GLM offrent une grande interprétabilité, où chaque coefficient β_j mesure l'effet marginal de la variable explicative correspondante. Ils sont bien adaptés aux relations linéaires ou log-linéaires, mais peuvent être limités pour modéliser des interactions complexes.

LightGBM, en revanche, est capable de capturer des relations non linéaires et des interactions complexes entre les variables explicatives. Cependant, il est moins interprétable et nécessite une attention particulière dans la sélection des hyperparamètres et le prétraitement des données.

Ces deux modèles, bien que très différents, offrent des perspectives complémentaires pour la tarification automobile. Les GLM garantissent un cadre transparent et conforme aux exigences réglementaires, tandis que LightGBM peut fournir une puissance prédictive supérieure dans des contextes plus complexes.

Bien que le but ici n'est pas la performance des modèles, il a été nécessaire de procéder à des évaluations, de manière à obtenir des modèles réalistes. Les modèles utilisés en pratique pour l'attribution des tarifs sont conçus pour être le plus performant possible. Se rapprocher de cet idéal a été l'un des objectifs pendant la modélisation. C'est pourquoi l'évaluation des modèles a été effectuée en utilisant des graphiques comme le **Lift-Chart** et des métriques comme le BIC (Critère d'Information Bayésienne) et surtout la **déviance expliquée** .

Évaluation des modèles de régression : déviance

La déviance : La déviance est une mesure clé de l'ajustement d'un modèle de régression, particulièrement dans les modèles linéaires généralisés. Elle quantifie l'écart entre les valeurs observées et celles prédites par le modèle, en comparant la vraisemblance du modèle ajusté à celle du modèle saturé (où chaque observation est parfaitement prédite). Pour une régression Poisson ($p = 1$), la déviance est donnée par :

$$D = 2 \sum_i \left(y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + \hat{y}_i - y_i \right).$$

De manière générale, la déviance peut s'exprimer comme :

$$D = 2 \cdot (\text{log-vraisemblance saturée} - \text{log-vraisemblance du modèle}),$$

où une déviance plus faible indique un meilleur ajustement du modèle.
La déviance expliquée : La déviance expliquée est une généralisation du R^2 pour les modèles de régression généralisés. Elle permet d'évaluer la proportion de variabilité des données expliquée par le modèle et se calcule comme suit :

$$\text{Déviance expliquée} = 1 - \frac{D_{\text{modèle}}}{D_{\text{nul}}},$$

où $D_{\text{modèle}}$ est la déviance du modèle ajusté, et D_{nul} est celle du modèle nul (modèle avec uniquement une constante). Une déviance expliquée proche de 1 reflète un bon pouvoir explicatif du modèle, tandis qu'une valeur proche de 0 indique un faible ajustement.
L'indice de Gini en régression : Dans le contexte des régressions, l'indice de Gini mesure l'inégalité des prédictions du modèle par rapport aux observations. Il est couramment utilisé pour évaluer la capacité du modèle à différencier les individus à haut risque des individus à faible risque. Sa formule s'exprime comme suit :

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |\hat{y}_i - \hat{y}_j|}{2n^2 \bar{\hat{y}}},$$

où $\hat{y}_i$ et $\hat{y}_j$ sont les valeurs prédites pour deux observations quelconques, n est la taille de l'échantillon, et $\bar{\hat{y}}$ est la moyenne des valeurs prédites. Un indice de Gini élevé reflète une forte capacité du modèle à distinguer les différents niveaux de risque.
Comparaison et avantages :
- La déviance et la déviance expliquée évaluent l'ajustement global du modèle et permettent de quantifier la variabilité expliquée par le modèle. Elles sont idéales pour comparer des modèles en termes de précision globale. - L'indice de Gini se concentre sur la dispersion et l'inégalité des prédictions, ce qui est particulièrement pertinent dans des applications où la différenciation des individus est essentielle, comme en tarification en assurance. Contrairement à la déviance, il ne dépend pas directement des valeurs observées, mais des différences relatives entre les prédictions.
Ces deux métriques sont donc complémentaires pour une évaluation approfondie des modèles de régression.

Après cette entrée en matière sur le fonctionnement des modèles, regardons de près leurs performances et surtout leurs niveaux d'équité. Rappelons que d'après la section4.2, les donnnées semblent ne pas présenter de différence car en dehors des situations extrêmes, hommes et femmes présentent des fréquences et sévérités de sinistres similaires.

4.4.2 Performances des modèles retenus

La modélisation de la fréquence et de la sévérité a été réalisée à l'aide de deux types de modèles : les modèles linéaires généralisés (GLM) et les modèles LightGBM (GBM). Chaque modèle a été optimisé via la validation croisée k-fold.

Avant cette validation, les données ont été scindées en un **échantillon d'entraînement** (80%) et un **échantillon de test** (20%) afin d'évaluer la capacité de généralisation des modèles. Cette séparation a été effectuée en **stratifiant par genre** pour préserver un équilibre hommes-

femmes et éviter un *biais de sélection* susceptible d'altérer les performances des modèles.

La validation croisée consiste à diviser l'échantillon d'entraînement en k sous-ensembles (*folds*). À chaque itération, un sous-ensemble est utilisé comme ensemble de validation, tandis que les $k-1$ autres servent à l'entraînement. Chaque sous-ensemble est ainsi utilisé une fois pour la validation, garantissant une évaluation robuste des performances et une meilleure généralisation en limitant les biais liés à une partition arbitraire des données. Les hyperparamètres sont optimisés en maximisant des métriques telles que la déviance expliquée moyenne et les résultats issus des Lift-Charts.

Les modèles retenus, c'est-à-dire ceux offrant les meilleures performances selon ces critères, ont ensuite été entraînés dans deux configurations distinctes :

— **Configuration « best-estimate »** : dans cette configuration, la variable sensible `DRIVER_GENDER` est incluse parmi les variables explicatives. Elle permet d'obtenir des résultats optimaux en termes de prédiction mais est utilisée uniquement à des fins d'analyse. Cette configuration n'est pas conforme à la réglementation en vigueur, notamment à la Directive Genre, et ne peut donc pas être utilisée pour la tarification effective.

— **Configuration « ignorante » (unaware)** : dans cette configuration, la variable sensible `DRIVER_GENDER` est exclue des variables explicatives. Cela garantit la conformité avec la Directive Genre, qui interdit l'utilisation des variables liées au genre dans la tarification des primes d'assurance.

Le tableau suivant présente les niveaux de déviance expliquée et les indices de Gini obtenus pour les différentes configurations de modèles et méthodologies.

Table 4.3 – Comparaison des métriques pour GBM et GLM

	Coût		**Fréquence**	
Modèle	**Témoin**	**Réglemenatire**	**Témoin**	**Réglementaire**
GBM				
Indice de Gini	6,03%	6,69%	22,50%	22,21%
Déviance expliquée	0,50%	0,43%	3,00%	2,88%
GLM				
Indice de Gini	10,19%	10,36%	22,28%	22,36%
Déviance expliquée	0,71%	0,83%	3,00%	2,98%

Comparaison des configurations *best-estimate* et *unaware*

En analysant les résultats, les écarts de performance entre les configurations *best-estimate* (incluant la variable sensible `DRIVER_GENDER`) et *unaware* (excluant cette variable) sont très faibles. Par exemple : - Pour le GBM appliqué aux coûts, la déviance expliquée est légèrement meilleure en configuration *best-estimate* (0,4960% contre 0,4255%). - Pour les fréquences, la différence entre les configurations est presque négligeable : par exemple, pour le GLM, la déviance expliquée passe de 2,98% (*unaware*) à 2,95% (*best-estimate*).

Ces résultats montrent que l'exclusion de la variable sensible n'affecte pas significativement les performances globales des modèles, ce qui peut suggérer que `DRIVER_GENDER` a un faible pouvoir explicatif dans ces données.

Comparaison entre les modèles GLM et GBM

De manière notable, le GLM surpasse systématiquement le GBM en termes de déviance expliquée pour toutes les combinaisons de configuration et de cible. Par exemple : - Pour les

coûts, la déviance expliquée atteint 0,007131 pour le GLM en configuration *best-estimate*, contre seulement 0,004960 pour le GBM. - Pour les fréquences, le GLM affiche des performances similaires ou légèrement supérieures au GBM, avec des déviances expliquées proches mais toujours un peu plus élevées (par exemple, 0,029524 contre 0,030035 en configuration *best-estimate*).

Les indices de Gini, bien qu'utiles pour mesurer la capacité de classement, n'inversent pas cette tendance. Même si le GBM est légèrement supérieur en termes de Gini pour certains cas (notamment sur les fréquences en configuration *best-estimate*), l'avantage est marginal comparé aux différences observées sur la déviance expliquée.

Le GLM, modèle linéaire par construction, semble mieux adapté dans ce cas précis, ce qui pourrait s'expliquer par plusieurs facteurs dont :

— 1. **l'absence de relations complexes dans les données** : Si les relations entre les variables explicatives et la cible sont principalement linéaires, le GBM, qui est conçu pour capturer des relations non linéaires et des interactions complexes, peut ne pas offrir d'avantage significatif .

— 2. **Le risque de surapprentissage pour le GBM** : Les GBM peuvent être sensibles au surapprentissage, particulièrement sur des ensembles de données où les relations complexes sont faibles.

Conclusion sur les choix de modèles Ces observations conduisent aux conclusions suivantes :

1. **Pour les coûts**, le GLM est clairement plus performant que le GBM en termes de déviance expliquée, même si le GBM peut légèrement surpasser le GLM en termes d'indice de Gini.

2. **Pour les fréquences**, les différences entre les deux modèles sont plus faibles, mais le GLM conserve un avantage, probablement grâce à sa capacité à mieux généraliser sur ces données.

Ainsi, pour l'évaluation de l'équité, le GLM est retenu à la fois pour la modélisation des coûts et des fréquences, en raison de sa performance globale supérieure et de sa robustesse face à des relations simples entre les variables. Le critère d'équité retenu reste l'indépendance, car il est simple à interpréter et permet une analyse comparative efficace entre les configurations *best-estimate* et *unaware*.

4.4.3 Critère d'indépendance sur les modèles de coût et de fréquence

À première vue, les modèles *Best-estimate* révèlent une différence notable dans les prédictions entre hommes et femmes. Cette asymétrie est visible dans la superposition des distributions des prédictions, marquant une disparité entre les deux sous-populations.

Cas des modèles de prédiction de coûts

Comme illustré dans le graphique ci-dessous, la disparité entre les prédictions des deux configurations est particulièrement marquée. Le modèle utilisant explicitement la variable sensible (`DRIVER_GENDER`) accentue l'asymétrie dans les distributions, avec des coûts moyens prédits plus élevés pour les hommes que pour les femmes. Ce constat est cohérent avec les analyses exploratoires mais présente une intensité amplifiée. Cela démontre que l'inclusion de la variable sensible induit un biais significatif dans le modèle, amplifiant les différences de risque perçues.

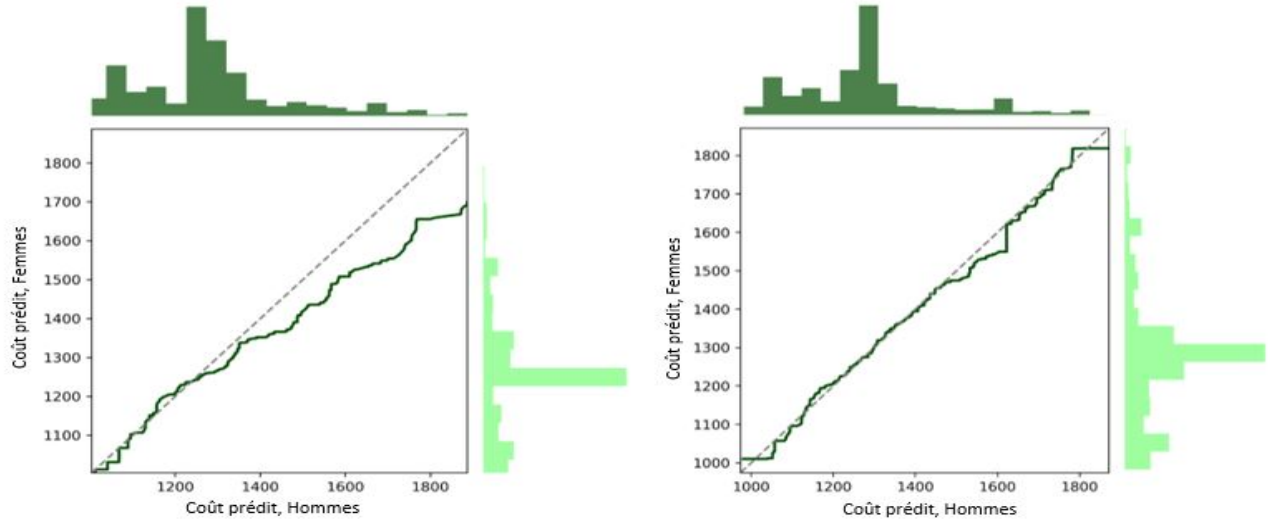

FIGURE 4.10 – Comparaison des distributions de coûts moyens prédits de sinistres selon le genre du conducteur. À gauche : prédictions du modèle *Best-estimate*. À droite : celles du modèle *Unaware*. L'asymétrie est plus prononcée dans la première configuration.

Cas des modèles de prédiction de fréquences

Les différences observées dans les prédictions de fréquence sont plus subtiles que pour les coûts. Globalement, comme dans les données exploratoires, les femmes présentent des fréquences de sinistres légèrement plus élevées, comme l'indique la courbe légèrement au-dessus de la première bissectrice au centre des distributions. Cependant, pour le modèle *Best-estimate*, cet ordre s'inverse dans les queues de distribution, où les hommes présentent des fréquences plus élevées. En revanche, la configuration *Unaware* tend à rétablir cette asymétrie en faveur d'une meilleure adéquation avec les observations empiriques.

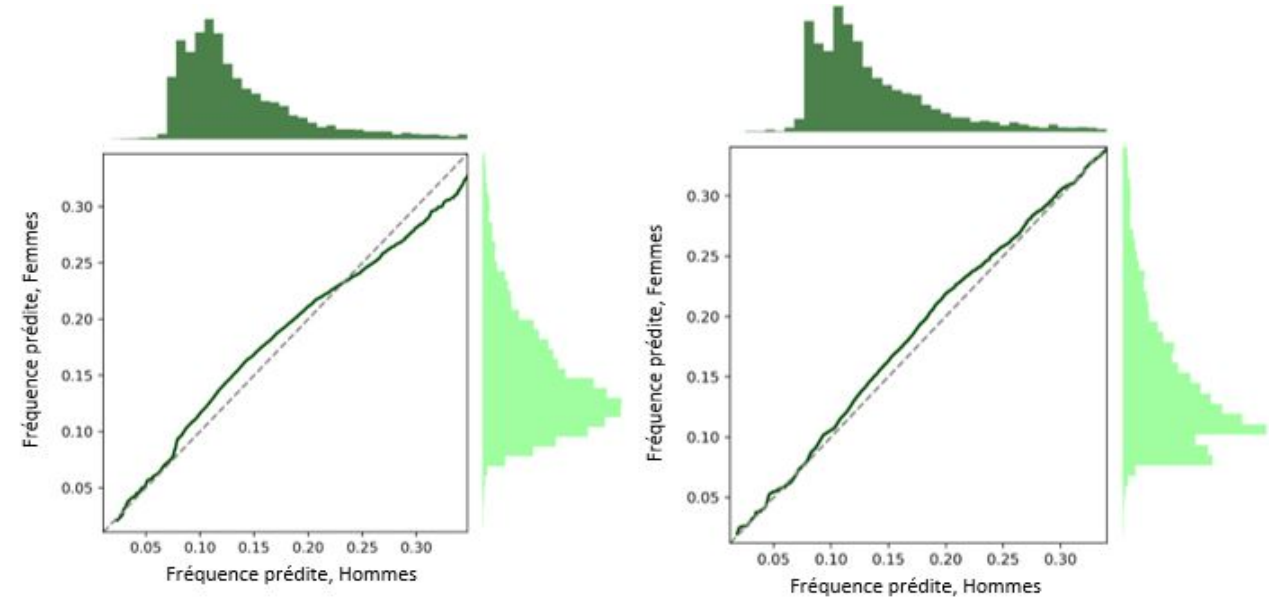

FIGURE 4.11 – Comparaison des distributions de fréquences annuelles prédites de sinistres selon le genre du conducteur À gauche : prédictions du modèle *Best-estimate*. À droite : celles du modèle *Unaware*. Les différences entre hommes et femmes sont moins marquées.

Examen des métriques d'équité

L'analyse des métriques confirme que l'indépendance des prédictions vis-à-vis de la variable sensible est mieux respectée dans les modèles *Unaware*. Comme illustré dans le tableau **??**, les modèles *Unaware* réduisent les écarts mesurés par les métriques d'équité telles que la distance de Kolmogorov (*Kolm-metric*) ou l'écart absolu moyen (*SPD*), tant pour la fréquence que pour le coût ou la prime pure. Par exemple, la différence moyenne de prime pure entre hommes et femmes diminue de 7,47% dans la configuration *Best-estimate* à 6,35% dans la configuration *Unaware*. Une tendance similaire est observée pour l'indicateur de la part de variance expliquée par la variable sensible, (*UF_metric*).

Cependant, ces améliorations demeurent insuffisantes pour atteindre une stricte parité démographique : les p-valeurs des tests de Kolmogorov et certains tests de Welch restent faibles, rejetant l'hypothèse d'indépendance parfaite.

TABLE 4.4 – Métriques des modèles « Best-estimate »

Métriques	**Fréquences**	**Coûts**	**Prime Pure**
Kolm-metric	$0,1473$	$0,2751$	$0,1598$
SPD (%)	$11,09$	$7,2124$	$7,4702$
UF_metric (%)	$2,1038$	$6,2272$	$0,5310$
P-valeur (Kolmogorov)	$0,0000$	$0,0000$	$0,0000$
P-valeur (Welch)	$0,0000$	$0,0000$	$0,0000$

TABLE 4.5 – Métriques des modèles « Unaware »

Métriques	**Fréquences**	**Coûts**	**Prime Pure**
Kolm-metric	$0,1030$	$0,0550$	$0,07963$
SPD (%)	$6,3522$	$0,0426$	$6,3532$
UF_metric (%)	$0,07078$	$0,0255$	$0,4113$
P-valeur (Kolmogorov)	$0,0000$	$0,00285$	$0,0000$
P-valeur (Welch)	$0,0000$	$0,2854$	$0,0000$

Ces résultats mettent en lumière que la suppression de la variable sensible peut atténuer les biais apparents et favoriser une meilleure équité des modèles. Toutefois, cette stratégie ne garantit pas une stricte indépendance statistique. Sur le plan business, une question clé persiste : ces différences résiduelles sont-elles acceptables ? Par exemple, les modèles classiques prédisent une prime pure moyenne de 190€ pour les femmes contre 180€ pour les hommes. Ces écarts, bien que statistiquement significatifs, pourraient ne pas être jugés discriminatoires si les différences observées dans le risque le justifient. La réponse à cette question dépendra du contexte économique, des exigences réglementaires et des attentes des parties prenantes, notamment des équipes commerciales et de leur analyse des compromis entre équité et performance actuarielle.
.

Chapitre 5

Approches pratiques pour la détection et la mitigation des biais en assurance automobile

Le chapitre précédent a permis de mesurer le niveau d'équité des modèles de tarification en assurance automobile en analysant les prédictions issues de différentes configurations. Bien que le retrait explicite de la variable sensible `DRIVER_GENDER` ait entraîné une diminution des écarts observés, il n'a pas permis d'atteindre une parité démographique, qu'elle soit stricte ou faible, d'un point de vue statistique. Ces résultats soulèvent une question fondamentale : quels sont les facteurs sous-jacents, potentiellement liés au genre, qui influencent les prédictions des modèles et participent au maintien des inégalités ? Ce cinquième chapitre vise à répondre à cette question en explorant les proxys du genre et en proposant des méthodes pour atténuer les biais identifiés.

Dans un premier temps, une analyse approfondie des variables explicatives sera menée afin d'identifier les proxys potentiels de la variable `DRIVER_GENDER`. Cette étape s'appuiera sur des méthodes d'importance des variables et des analyses multivariées pour déterminer dans quelle mesure certaines variables influencent indirectement les prédictions et contribuent aux déséquilibres observés. En complément, les métriques d'équité seront réexaminées afin de mieux quantifier l'impact de ces proxys sur les résultats des modèles.

Par la suite, des approches pratiques de mitigation seront explorées et appliquées. Ces méthodes se déclineront en deux volets : des ajustements réalisés avant la modélisation, visant à réduire l'influence des proxys identifiés dans les données, et des stratégies post-modélisation, ayant pour objectif de corriger directement les prédictions afin de rétablir une équité relative. Les performances et les limites de ces approches seront analysées dans le contexte de la tarification automobile.

Enfin, une réflexion plus large sera menée pour discuter des implications des résultats obtenus, en tenant compte des enjeux réglementaires, techniques et commerciaux propres au secteur de l'assurance. Des recommandations seront proposées pour guider la mise en œuvre d'une tarification plus équitable.

Ce chapitre est organisé comme suit :

— La Section 5.1.2 identifie les proxys du genre à travers une analyse des variables explicatives et des métriques d'équité.

— La Section 5.2 applique des méthodes de mitigation avant et après modélisation, en évaluant leur efficacité et leurs limites.

— La Section 5.3 discute des résultats obtenus et propose des recommandations pratiques pour intégrer l'équité dans les processus de tarification.

5.1 Détection des causes de biais

L'objectif de cette section est d'identifier les variables non sensibles qui contribuent indirectement aux disparités observées dans les modèles, en jouant le rôle de proxys pour la variable sensible `DRIVER_GENDER`. Dans un premier temps, une analyse de l'importance des variables explicatives sera réalisée pour chacun des modèles sélectionnés, en particulier les modèles *GLM* étudiés dans le chapitre précédent. Cette analyse permettra de mettre en évidence les variables ayant une influence significative sur les prédictions.

Dans un second temps, l'impact de ces variables sur les disparités observées sera examiné à l'aide des métriques d'équité. En croisant les résultats sur l'importance des variables et leur contribution aux déséquilibres, il sera possible d'identifier les proxys potentiels du genre. Ces informations constitueront une base essentielle pour les étapes ultérieures de mitigation des biais.

5.1.1 Importance des variables

Pour évaluer l'importance des variables dans les modèles prédictifs, la **valeur de Shapley** est un outil particulièrement adapté. Issue de la théorie des jeux coopératifs, elle permet de mesurer la contribution individuelle de chaque variable explicative aux prédictions d'un modèle, tout en prenant en compte l'interaction entre les variables. L'idée centrale est de répartir de manière équitable l'impact total d'une prédiction parmi toutes les variables en fonction de leur rôle spécifique dans le modèle.

En termes simples, la valeur de Shapley évalue la manière dont chaque variable influe sur les prédictions en comparant les résultats obtenus avec et sans cette variable. Cette analyse est réalisée pour chaque instance individuelle des données, ce qui permet de calculer des contributions spécifiques à chaque observation et de détecter les variations locales des contributions des variables explicatives. Cela rend la méthode particulièrement utile dans le contexte de modèles complexes ou non linéaires.

Sur le plan technique, pour une variable donnée x_i et une instance individuelle x, la valeur de Shapley $\phi_i(x)$ est définie par la formule suivante :

$$\phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left(f(S \cup \{i\}) - f(S)\right), \text{ où :}$$

— N est l'ensemble des variables explicatives,

— S représente un sous-ensemble des variables ne contenant pas x_i,

— $f(S)$ désigne la sortie du modèle lorsqu'il est construit uniquement à partir des variables de S,

— $f(S \cup \{i\})$ est la sortie du modèle après ajout de la variable x_i au sous-ensemble S.

Cette formule s'appuie sur un calcul combinatoire pour évaluer toutes les façons possibles d'ajouter x_i aux différentes combinaisons de variables. Ces valeurs sont calculées pour chaque instance individuelle des données, permettant d'attribuer des contributions uniques à chaque variable en fonction de son rôle spécifique dans la prédiction pour cette observation.

En pratique, le calcul direct des valeurs de Shapley est souvent prohibitif en termes de temps de calcul, car le nombre de combinaisons possibles croît de manière exponentielle avec le nombre de variables. Pour pallier cette difficulté, des méthodes d'estimation basées sur des échantillonnages aléatoires (*Monte Carlo sampling*) sont utilisées. Ces approches permettent d'obtenir une approximation des valeurs de Shapley en réduisant considérablement la charge computationnelle tout en maintenant une précision acceptable.

Dans le cadre d'une régression, les valeurs de Shapley s'interprètent comme des contributions additives expliquant la prédiction individuelle du modèle. Pour une instance donnée x, la prédiction du modèle peut être décomposée comme suit :

$$f(x) = \text{valeur\_de\_base} + \sum_{i=1}^{n} \phi_i(x), \text{ où :}$$

— $f(x)$ est la prédiction pour l'instance x,

— valeur_de_base représente la valeur moyenne des prédictions du modèle sur l'ensemble des données (ou une autre base de référence),

— $\phi_i(x)$ est la valeur de Shapley de la variable x_i pour l'instance x, indiquant sa contribution à la différence entre la prédiction individuelle et la valeur de base.

Cette décomposition additive rend l'interprétation intuitive et flexible. Elle permet de comprendre, à l'échelle individuelle, comment chaque variable influence une prédiction donnée. Les valeurs de Shapley offrent ainsi une vision globale de l'importance des variables dans un modèle, tout en fournissant des insights précis et localisés pour chaque observation analysée.

Interprétation des modèles de sévérité

Les valeurs de Shapley (Shap) obtenues sur les modèles GLM en version « Best-estimate » et en version « Unaware » fournissent des informations riches et détaillées, comme illustré dans le diagramme en essaim ci-dessous. Chaque point du diagramme représente une valeur de Shapley individuelle associée à une observation de l'échantillon test, pour chacune des variables explicatives.

Premièrement, l'ordre des variables sur l'axe vertical indique leur importance relative dans les prédictions. On observe que la variable `coverage` (type de couverture) est la plus influente dans la prédiction des coûts. Cela semble intuitif, car un niveau de couverture plus élevé implique une prise en charge de risques plus importants en cas de sinistre. La deuxième variable majeure est l'expérience de conduite, mesurée par la notation bonus-malus (`bm`). Les assurés avec un historique dégradé (notation élevée) tendent à engendrer des sinistres plus coûteux. Cet effet est asymétrique : les assurés avec une mauvaise notation présentent des coûts nettement plus élevés, tandis qu'une bonne notation entraîne des réductions moins marquées des coûts. Cependant, cette variable est souvent indisponible lors de la tarification initiale, car elle repose sur l'historique des contrats précédents.

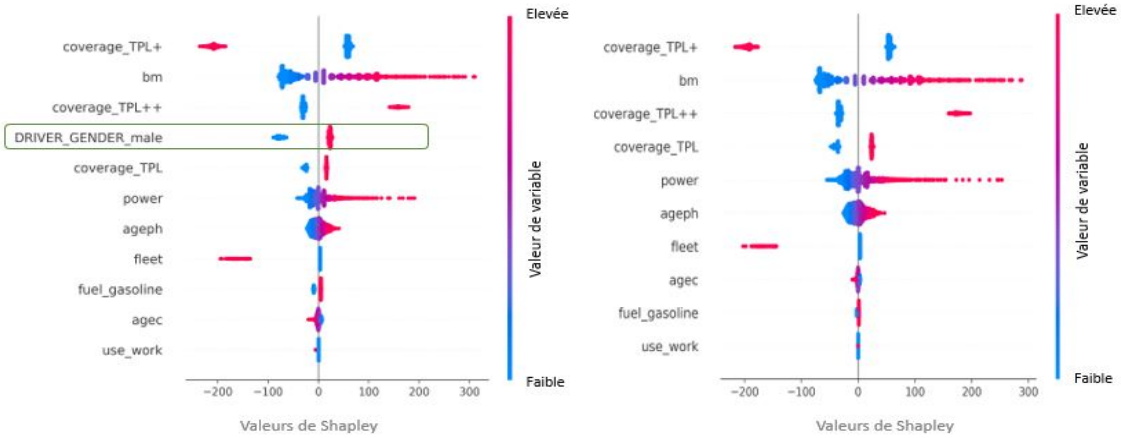

FIGURE 5.1 – Diagramme en essaim des valeurs de Shapley pour les modèles de coût. À gauche : prédictions du modèle *Best-estimate*. À droite : celles du modèle *ignorant*.

Dans le modèle « Best-estimate », la troisième variable la plus influente est le genre du conducteur, une variable sensible. Les diagrammes révèlent que les femmes présentent généralement un coût de sinistre prédit inférieur d'au moins 50 € par rapport aux hommes, toutes choses égales par ailleurs. Cela constitue une forme de ***discrimination directe***, car le genre est explicitement pris en compte dans les prédictions. De plus, cette variable est statistiquement significative, renforçant son poids dans les prédictions.

Lorsque la variable "genre" est exclue (« modèle ignorant »), cette discrimination directe disparaît. Dans ce cas, la troisième variable la plus influente devient la puissance du véhicule (`power`). Une puissance plus élevée entraîne des coûts de sinistres plus importants, probablement parce que les véhicules puissants sont souvent plus coûteux à réparer et causent davantage de dégâts aux tiers. Ces véhicules sont également susceptibles d'être plus rapides ou plus lourds, amplifiant les impacts (et donc les coûts) en cas d'accident. D'autres variables, comme l'âge du conducteur et du véhicule, l'appartenance à une flotte, le type d'usage, ou le carburant utilisé, ont des effets marginaux dans ce modèle.

Interprétation des modèles de fréquences

En appliquant le même cadre d'analyse, les valeurs de Shapley permettent d'identifier les variables les plus influentes pour les prédictions de fréquence. Dans l'ordre décroissant d'importance, les variables majeures sont : le score de conduite (`bm`), l'âge du conducteur, le type de carburant, la puissance du véhicule, et le type de couverture. Dans ce modèle, le genre du conducteur n'apparaît qu'en sixième position, avec un impact marginal, même dans le modèle « Best-estimate ». Ainsi, l'introduction explicite de cette variable sensible n'ajoute pas de biais supplémentaire dans les prédictions de fréquence.

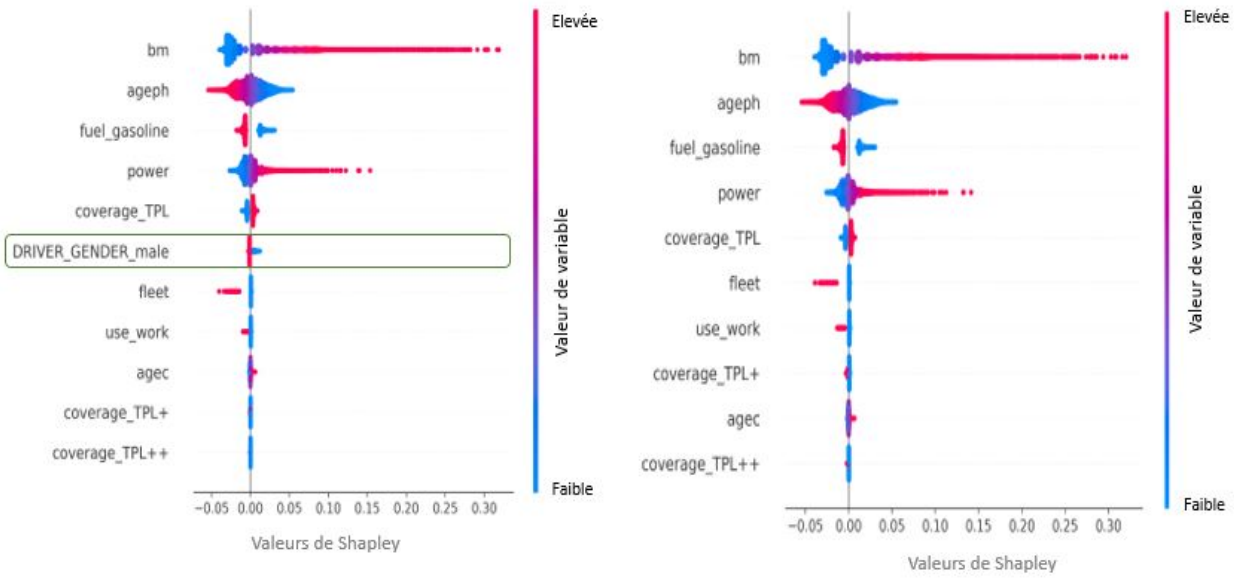

FIGURE 5.2 – Diagramme en essaim des valeurs de Shapley pour les modèles de fréquence. À gauche : prédictions du modèle *Best-estimate*. À droite : celles du modèle *ignorant*.

En analysant plus en détail les prédictions, on observe des relations intéressantes :

— Les assurés plus âgés tendent à présenter des fréquences de sinistres plus faibles, toutes choses égales par ailleurs.

— Les véhicules utilisant du diesel ont des sinistres plus fréquents que ceux alimentés par d'autres types de carburant. Cela pourrait s'expliquer par le fait que les véhicules diesel sont souvent utilisés pour des trajets plus longs ou plus fréquents.

— Les véhicules puissants sont associés à une fréquence de sinistres plus élevée, probablement en raison de comportements de conduite plus risqués ou d'une utilisation plus intensive.

D'autres variables, comme l'appartenance à une flotte (`fleet`), le type d'usage (`use`), ou l'âge du véhicule, n'ont pas d'impact significatif sur les prédictions de fréquence.

Synthèse des observations

Il ressort de cette analyse que, même lorsqu'il est inclus, le genre du conducteur reste une variable secondaire dans les modèles de fréquence (sixième position), alors qu'il est la troisième variable la plus influente dans les modèles de coût. Cela montre que la discrimination directe est plus marquée dans le modèle « Best-estimate » de sévérité que dans celui de fréquence. Par construction, cette discrimination disparaît dans les modèles « Unaware ». Cependant, dans ces derniers, on observe encore des disparités de prédictions entre hommes et femmes. Cette question soulève l'intérêt d'examiner plus en profondeur les proxys indirects et les relations cachées entre variables. Ceci ayant déjà fait l'objet de la section 4.3.2, il est question ici d'examiner de façon concrète les proxy au travers des métriques d'équité.

5.1.2 Analyse des métriques d'équité par permutation

Dans les modèles ignorants, bien que la discrimination directe soit éliminée, des disparités subsistent entre les prédictions pour les hommes et les femmes (bien que faibles au sens business). Ces différences soulèvent l'intérêt d'examiner les proxys indirects responsables de ces disparités. Alors que la Section 4.3.2 s'est concentrée sur une analyse observationnelle, reposant sur les corrélations entre les variables explicatives, la cible, et la variable sensible sans recours aux modèles, cette sous-section adopte une approche centrée sur les modèles pour identifier les proxys.

L'objectif ici est de mesurer la contribution spécifique de chaque variable explicative aux niveaux d'inéquité dans les modèles. Pour ce faire, une technique simple mais puissante issue du machine learning est utilisée : la **méthode d'importance par permutation** (ou *permutation importance*).

Concept et définition. L'importance par permutation est une méthode permettant d'évaluer l'influence d'une variable explicative en mesurant l'impact de la perturbation de cette variable sur une métrique cible. Dans le contexte de l'équité, cette technique évalue comment la perturbation d'une variable affecte les métriques d'équité d'un modèle.

Concrètement, pour une variable donnée X_j, les valeurs observées dans la colonne correspondante de l'ensemble des données X sont permutées de manière aléatoire, tout en gardant les autres variables inchangées. Cette opération brise la relation entre x_j et les autres variables ou la cible y. Le modèle est alors réévalué en utilisant cet ensemble permuté, et les métriques d'équité sont recalculées. L'importance de la variable x_j est estimée comme la différence moyenne entre la valeur de la métrique d'équité sur l'ensemble original et sur les ensembles permutés.

Formule de calcul. Soit M le modèle, X l'ensemble des données explicatives, y la cible, et $\mathcal{F}$ une métrique d'équité (comme la *statistical parity difference* `SPD` ou la distance de Kolmogorov `Kolm`). L'importance par permutation pour une variable x_j est donnée par :

$$I(x_j) = \frac{1}{N} \sum_{k=1}^{N} \left| \mathcal{F}(M(X, y)) - \mathcal{F}(M(X^{\text{permute}(j)}, y)) \right|, \text{ où :}$$

— $X^{\text{permute}(j)}$ correspond à l'ensemble des données où la colonne x_j a été permutée aléatoirement,

— N est le nombre de répétitions de la permutation.

Cette méthode s'appuie sur la répétition (*n_repeats*) pour garantir la robustesse des estimations, car une seule permutation pourrait ne pas capturer l'impact réel de x_j.

Application dans le cadre de l'équité. La méthode d'importance par permutation a été appliquée aux deux métriques principales d'équité utilisées dans cette étude :

— `SPD` (*Statistical Parity Difference*), mesurant l'indépendance au sens faible entre la variable sensible `DRIVER_GENDER` et les prédictions.

— `Kolm` (distance de Kolmogorov), qui évalue l'indépendance au sens strict en comparant les distributions des prédictions pour les hommes et les femmes.

Le processus consiste à perturber successivement chaque variable explicative, puis à observer les variations dans les valeurs de `SPD` et `Kolm`. Les variables ayant le plus grand impact sur ces métriques sont identifiées comme des proxys potentiels de la variable sensible.

Interprétation des résultats. Cette méthode offre une perspective quantitative claire sur l'influence des variables explicatives. Les variables présentant une importance élevée pour les métriques d'équité sont suspectées d'être des proxys, car leur perturbation modifie significativement les niveaux d'inéquité. Par exemple, une forte variation dans `SPD` après la permutation d'une variable pourrait indiquer une relation indirecte entre cette variable et la variable sensible `DRIVER_GENDER`.

Avantages et limites. L'importance par permutation présente plusieurs avantages :

— Elle est facile à implémenter et indépendante du type de modèle.

— Elle quantifie explicitement l'impact des variables sur une métrique d'intérêt.

Cependant, elle présente aussi des limitations, notamment :

— Une charge computationnelle importante pour les modèles complexes ou les ensembles de données volumineux.

— Une sensibilité aux corrélations entre variables, qui peut biaiser les estimations d'importance.

Analyse des proxys dans les modèles de coût

L'analyse des proxys dans les modèles de coût, réalisée à l'aide des métriques de parité démographique faible (SPD) et stricte (Kolmogorov), met en lumière plusieurs dynamiques importantes. Ces observations révèlent des différences significatives entre les configurations *best-estimate* et *unaware*, tout en soulignant l'impact de la suppression de la variable sensible `DRIVER_GENDER` sur l'importance des autres variables.

Observations sur le modèle *best-estimate* Dans la configuration *best-estimate*, où la variable `DRIVER_GENDER` est incluse comme explicative, cette dernière occupe une position importante, bien qu'elle ne soit pas systématiquement la variable la plus influente. Pour le SPD, sa contribution est de 0.0197, la plaçant derrière `ageph` (0.0352) et `agec` (0.0425). Pour Kolmogorov, elle occupe une place similaire avec une valeur de 0.1581, mais elle est devancée par des variables comme `ageph` (0.2371) et `agec` (0.2934).

Ces résultats s'expliquent par la corrélation entre ces variables explicatives et la variable sensible. Par exemple, `ageph`, `power` et `bm` sont historiquement liés au genre comme l'illustrent les graphiques de la figure 4.9 du chapitre 4. Il n'en est pas de même pour l'âge des véhicules `agec` qui historiquement est très peu liée au genre. Sa prédominance comme proxy dans les

modèles *best-estimate* peut s'expliquer par un effet d'interaction dans le modèle. Lorsque la variable `DRIVER_GENDER` est incluse, `agec` pourrait capter des informations indirectes sur les disparités de genre en raison de son rôle combinatoire avec d'autres variables explicatives. Par exemple, `agec` pourrait agir comme un ajustement contextuel qui, bien que non lié directement au genre ou aux coûts, amplifie les effets de variables historiquement biaisées comme `power` ou `bm` (d'autant plus qu'elle y est liée 4.5). Une autre explication réside dans la nature additive des modèles utilisés, notamment le GLM. Dans ce type de modèle, les interactions complexes ne sont pas directement modélisées, mais certaines variables peuvent compenser ou ajuster les biais générés par d'autres.

Cependant, la variable `DRIVER_GENDER`, bien qu'influente, ne domine pas complètement les prédictions, ce qui pourrait indiquer une interaction complexe entre les différentes variables explicatives.

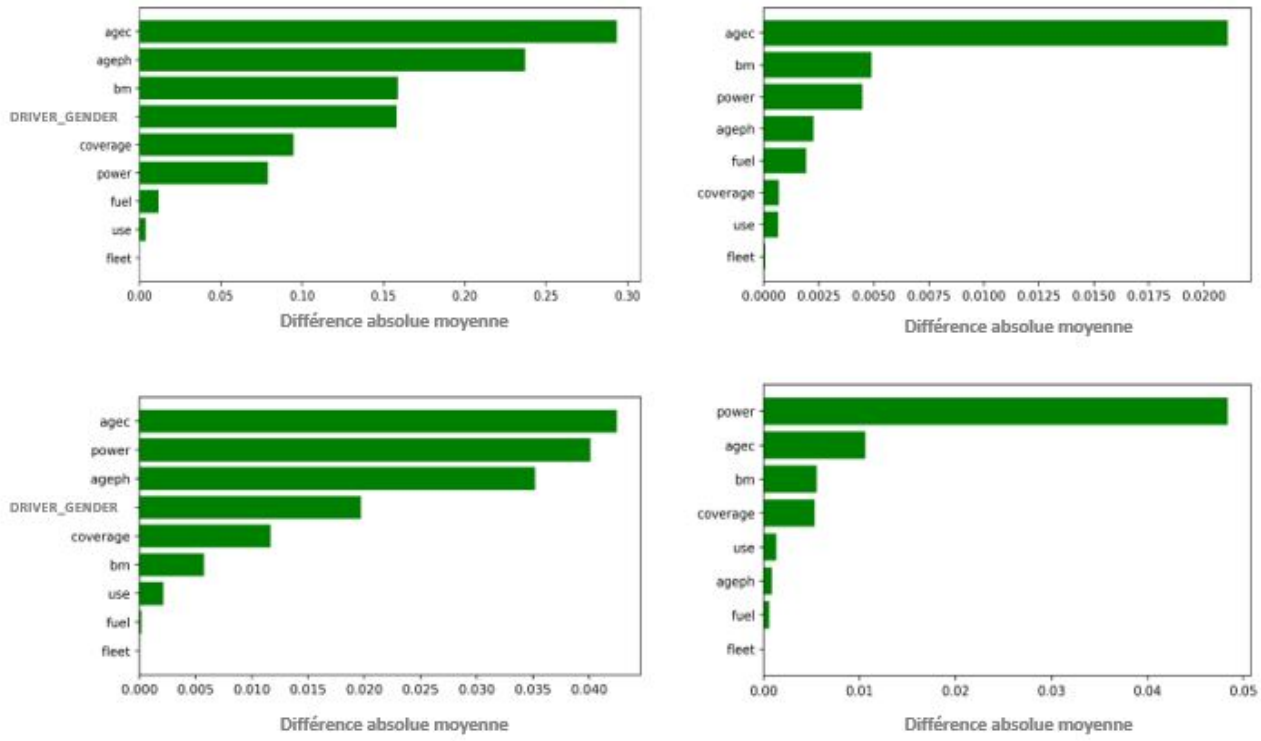

FIGURE 5.3 – Importance par permutation des variables pour les modèles de coût. À gauche : modèles *Best-estimate*. À droite : modèles *ignorants*. La $1^{ère}$ ligne concerne la distance de Kolmogorov et la seconde la SPD

Comparaison avec le modèle *unaware* Dans la configuration *unaware*, l'absence explicite de `DRIVER_GENDER` modifie significativement les proxys. Les variables comme `ageph` (0.0005 pour SPD et 0.0023 pour Kolmogorov) et `agec` (0.0103 pour SPD et 0.0211 pour Kolmogorov) prennent une part accrue dans les disparités résiduelles.

Un point important à souligner est l'ordre de grandeur des importances des variables : les valeurs absolues des différences moyennes d'impact sur les métriques d'équité sont considérablement réduites dans le modèle *unaware*. Par exemple, pour `agec`, l'importance passe de 0.0425 à 0.0103 pour le SPD et de 0.2934 à 0.0211 pour Kolmogorov. Cette diminution montre que, bien que des proxys persistent, leur capacité à transmettre des biais est largement atténuée lorsque `DRIVER_GENDER` est supprimée.

Comparaison entre SPD et Kolmogorov Les proxys identifiés diffèrent légèrement selon les métriques utilisées. Tandis que le SPD, basé sur les moyennes globales, met davantage en

avant des proxys tels que `power` (0.0486 en *unaware*) ou `bm` (0.0052 en *unaware*), la distance de Kolmogorov capte mieux les disparités dans les queues des distributions, où `agec` et `ageph` jouent un rôle prépondérant. Cela reflète leur influence structurelle sur les prédictions, même après suppression de `DRIVER_GENDER`.

Conclusion L'ordre de grandeur des proxys diminue de manière significative dans la configuration *unaware*, indiquant que l'exclusion explicite de `DRIVER_GENDER` réduit l'intensité globale des biais. Toutefois, les proxys tels que `ageph`, `agec`, et `power` continuent de transmettre une partie des disparités, en raison de leur forte corrélation avec la variable sensible.

Cette réduction des importances absolues dans le modèle *unaware* reflète une atténuation des biais, mais elle ne garantit pas leur élimination totale.

Analyse des proxys dans les modèles de fréquence

Les proxys identifiés dans les modèles de fréquence montrent des différences significatives entre les configurations *best-estimate* et *unaware*. Ces observations sont discutées séparément pour chaque métrique d'équité, à savoir la **Différence de Parité Statistique (SPD)** et la **Distance de Kolmogorov (Kolm)**.

Observations pour la métrique SPD Dans la configuration *best-estimate*, les trois premiers facteurs influents identifiés sont `DRIVER_GENDER` (0,0449), suivi de `bm` (0,0365) et `power` (0,0274). Ces résultats reflètent directement la relation entre ces variables et le genre, comme discuté dans les analyses du chapitre précédent. La variable `ageph` (0,0842) domine largement en raison de sa forte corrélation historique avec le genre et de son rôle prépondérant dans les prédictions de fréquence.

Dans la configuration *unaware*, l'ordre change de manière notable. `ageph` devient la variable la plus influente (0,0499), suivie de `bm` (0,0652) et de `agec` (0,0217). Cette réorganisation met en évidence un rééquilibrage dans l'importance des proxys après la suppression explicite de `DRIVER_GENDER`. Le rôle accru de `agec` peut être attribué à des interactions complexes, bien qu'elle ne soit ni directement liée au genre ni historiquement corrélée à la fréquence des sinistres.

Observations pour la métrique Kolmogorov (Kolm) Pour la configuration *best-estimate*, les proxys les plus influents sont `ageph` (0,0616), suivi de `DRIVER_GENDER` (0,1629) et `bm` (0,0847). Ici encore, la domination de `ageph` est attendue, mais le rôle prépondérant de `DRIVER_GENDER` indique une forte contribution directe de la variable sensible à la disparité mesurée par la distance de Kolmogorov.

Dans la configuration *unaware*, l'importance relative des proxys change. `bm` (0,0427) prend la première place, suivi de `ageph` (0,0571) et de `power` (0,0219). Cette redistribution est cohérente avec le fait que les proxys liés au comportement de conduite (tels que `bm` et `power`) compensent la perte de la variable sensible.

Comparaison des configurations *best-estimate* et *unaware* L'analyse révèle que les trois premiers proxys identifiés dans chaque configuration sont similaires pour les deux métriques d'équité (`ageph`, `bm`, `power`), bien que leur ordre diffère légèrement. Ce constat met en évidence que ces variables sont structurellement liées aux prédictions des modèles, même après suppression explicite de `DRIVER_GENDER`.

Un point notable est la diminution de l'ordre de grandeur des importances moyennes dans la configuration *unaware*. Par exemple, pour la métrique SPD, l'importance de `ageph` passe de 0,0842 (*best-estimate*) à 0,0499 (*unaware*), et celle de `bm` passe de 0,0365 à 0,0652. Cette

atténuation généralisée reflète la suppression directe du signal associé à la variable sensible, ce qui réduit l'intensité des disparités mesurées.

Conclusions sur les proxys des modèles de fréquences Les proxys dominants dans les modèles de fréquence (notamment `ageph`, `bm`, et `power`) restent cohérents entre les configurations *best-estimate* et *unaware*, bien que leurs importances relatives soient modifiées. La réduction significative des valeurs d'importance moyenne après suppression de `DRIVER_GENDER` souligne l'impact direct de la variable sensible sur les disparités mesurées. Néanmoins, des proxys indirects, tels que `agec`, continuent de jouer un rôle non négligeable, ce qui met en lumière la complexité des relations entre variables dans les modèles de fréquence.

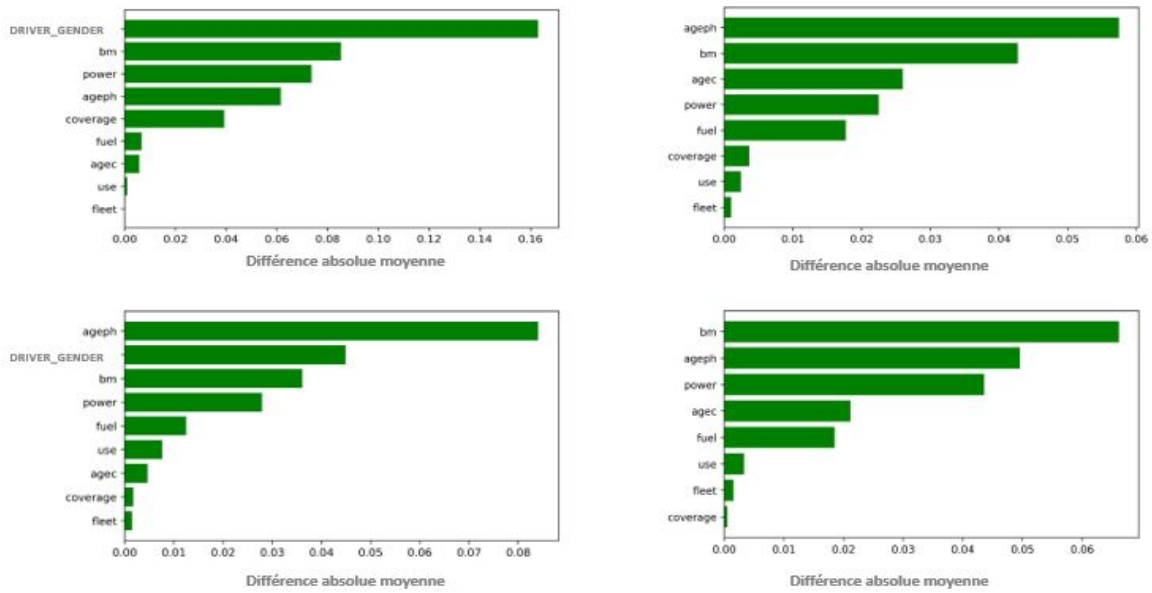

FIGURE 5.4 – Importance pour les modèles de fréquence par permutation des variables. À gauche : modèles *Best-estimate*. À droite : modèles *ignorants*. La $1^{ère}$ ligne concerne la distance de Kolmogorov et la seconde la SPD

Après cette analyse fine des variables légitimes, il ressort que les principaux proxys dans les modèles ignorants incluent : l'âge du véhicule `agec`, sa puissance `power`, l'âge du conducteur `ageph` et son historique de sinistre `bm`. Ceci permet d'appliquer des techniques de mitigation (surtout pré-processing) de manière potentiellement plus ciblée.

5.2 Application des méthodes de mitigation

Rappelons que les primes pures prédites au chapitre 4 présentaient une disparité (au sens su principe d'indépendance) statistiquement significatives, bien que relativement faibles d'un point de vue business (différence moyenne de primes pure de l'ordre de 6%). Le but ici est d'utiliser les méthodes de mitigation pour réduire l'aspect statistique des disparités, et de fait améliorer l'équité d'un point de vue business. En raison de leurs relatives simplicités computationnelles, les méthodes pré-processing d'orthogonalisation et post-processing de transport optimal sont examinées.

5.2.1 Orthogonalisation des features non sensibles

Il s'agit de la méthode décrite dans la section 3.1.2. Le but est de décorreler les variables explicatives légitimes de la variables sensible. Le paramètre $\alpha \in [0; 1]$ permet de maîtriser le

niveau de décorrélation appliqué et donc la quantité d'information transformée et non interprétable de chaque variable incorporée dans le modèle. Le choix de ce paramètre doit être fait de manière méthodique, car il a le potentiel d'influer sur les performances et l'équité des modèles. L'implémentation est réalisée à l'aide de la librairie *Fairlean*[1] en langage Python.

Pour y parvenir, nous réalisons, sur la base d'entrainement un ensemble d'étapes pour le choix optimal du α. Dans un ensemble fini de valeurs possibles, l'orthogonalisation est effectuée, les modèles de fréquence et de coûts sont entraînés. Ils sont ensuite évalués en terme de pouvoir prédictif (via l'indice de Gini, la déviance expliquée) et de parité démographique stricte (par la distance de `Kolmogorov`) ou faible (par la différence de parité `SPD`).

Pour les modèles de fréquence : Le paramètre alpha affecte effectivement la qualité prédictive du modèle et même son niveau de parité démographique. Son augmentation entraine une amélioration de la parité (au sens strict comme faible), et une baisse de la performance. Ceci est visible au travers de la distance de Kolmogorov et l'indice de Gini représentés sur le graphique ci-dessous. Alors que l'indice de performance Gini diminue de $1,3\%$ seulement , celle d'équité s'améliore de 45% ; l'équité s'améliore plus rapidement que la performance se dégrade.

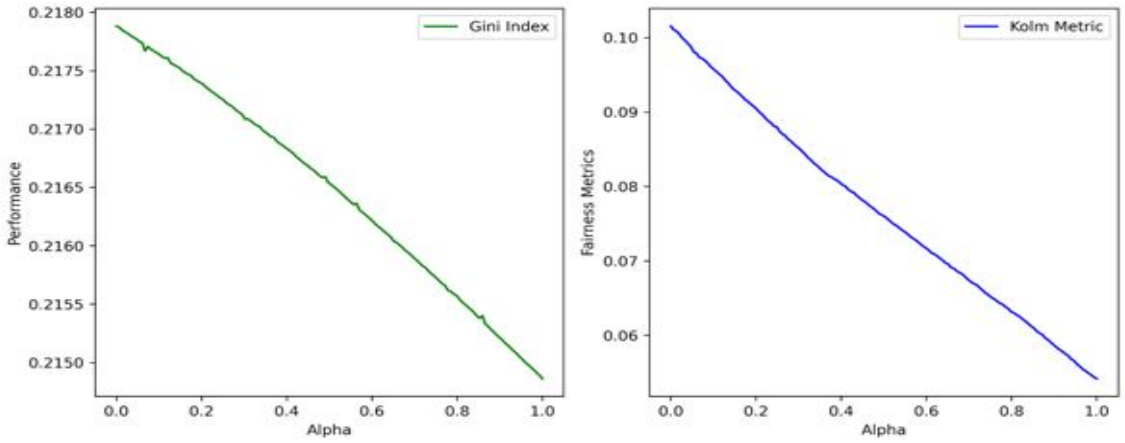

FIGURE 5.5 – Evolution du pouvoir prédictif et de l'équité du modèle de fréquences selon α.

L'exemple met en exergue le besoin d'arbitrage entre performance et équité, mais est-ce toujours le cas ?

Pour les modèles de coût : L'observation est différente et mérite une analyse approfondie. Contrairement aux attentes, l'évolution du niveau d'équité avec α n'est pas monotone. Que ce soit avec la distance de Kolmogorov (pour la mesure stricte) ou avec la différence relative de moyennes (`SPD`), on observe une amélioration de l'équité jusqu'à un certain seuil ($\alpha \in [0.7; 0.85]$), suivie d'une dégradation au-delà de ce niveau. Pendant ce temps, la performance s'améliore de manière quasi-continue avec α.

La dégradation de l'équité au-delà du seuil optimal ($\alpha \approx 0.77$) est contre-intuitive. En théorie, augmenter l'orthogonalisation des features X vis-à-vis du genre devrait réduire l'influence de la variable sensible et, par conséquent, améliorer l'équité. Cependant, ce phénomène peut s'expliquer par le fait qu'au-delà d'un certain niveau, une orthogonalisation excessive perturbe la structure des relations entre les variables explicatives et la cible. En forçant une indépendance trop stricte entre les features et le genre, certaines interactions pertinentes pour l'équité peuvent

1. https ://fairlearn.org/

être déformées ou perdues, créant ainsi des biais inattendus dans les prédictions. Cela se traduit par une augmentation de la distance de Kolmogorov de 53% après le seuil d'amélioration initial, suggérant que la structure des dépendances essentielles est compromise.

En ce qui concerne l'amélioration continue des performances avec α, bien que l'augmentation soit relativement modeste (1,4%), elle reste significative. Cela indique que la présence d'une composante liée au genre dans les variables explicatives est préjudiciable à la qualité prédictive du modèle. Historiquement, la sévérité des sinistres a été observée comme dépendante du genre (voir les p-valeurs du test de Kolmogorov 4.2). Par conséquent, un modèle dont les variables explicatives respectent ce principe est susceptible de mieux généraliser, ce qui explique l'amélioration des performances lorsque l'orthogonalisation est appliquée de manière appropriée.

Ces résultats mettent en évidence la nécessité de trouver un équilibre subtil dans l'orthogonalisation. Une approche trop rigide peut perturber la structure des données et nuire à l'équité, tandis qu'une orthogonalisation modérée peut contribuer simultanément à une meilleure équité et à une amélioration des performances.

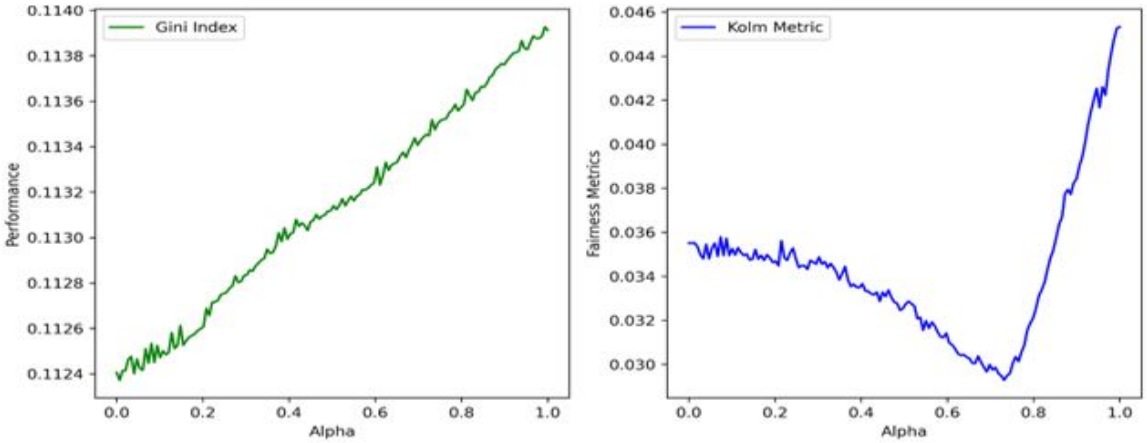

FIGURE 5.6 – Evolution du pouvoir prédictif et de l'équité du modèle de coûts selon α.

Parité démographique des Primes pures obtenues après orthogonalisation

En choisissant le niveau le plus optimal (en terme d'équité) d'orthogonalisation pour le modèle de coûts, c'est-à dire $\alpha^*_{cts} = 0,77$ et celui de fréquences $\alpha^*_{freq} = 1$, la base de données initiale est transformée en deux bases pour chaque modèle. Après leurs entrainements et leurs évaluations, les primes pures prédites sont calculées. Le résultat qu'on obtient est que le niveau de parité démographique (strict comme faible) est meilleur que celui du modèle « ignorant » (unaware) du chapitre précédent. A titre illustratif, la différence moyenne de prime pure passe de $6,35\%$ (voir) à $1,12\%$, et la distance de Kolmogorov de $0,08$ à $0,06$, une évolution beaucoup moins importante.

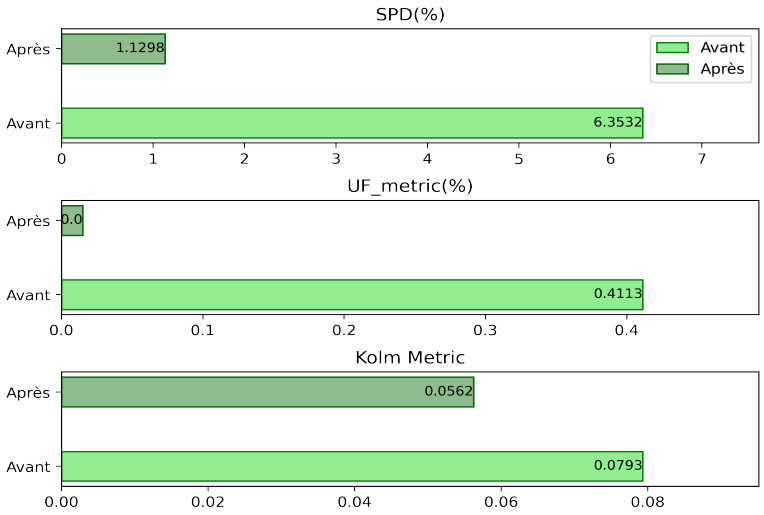

FIGURE 5.7 – Evolution des métriques d'équité de prime pure avant et après orthogonalisation

Cependant, la significativité statistique ne change pas véritablement, car les différences, bien que faibles demeurent significatives au seuil de 1%. Pour en avoir la certitude, le graphe de correspondance de prime pure obtenue après mitigation nous donne plus d'informations. On observe par exemple que pour les 75% plus faibles primes, les femmes paient un peu plus que les hommes. Et pour ce qui est des 20% plus élevées, les primes sont un peu plus élevées chez les hommes. Pourtant, avant mitigation, le niveau de prime pure était globalement plus élevé chez les hommes. De plus, en comparant les primes pures initiales et après mitigation, aucune tendance globale ne se dégage. On retrouve des changements qui peuvent être très importants (voir figure 5.15 en annexe). Ceci met donc à mal l'implémentation de cette méthode d'un point de vue marketing et explication des révisions tarifaires.

La significativité des tests peut être due soit à la taille importante de l'échantillon utilisé pour le test (20% de la base de données) ou à la qualité même de la méthode. La technique du transport optimal post-modélisation permet-elle de faire mieux ?

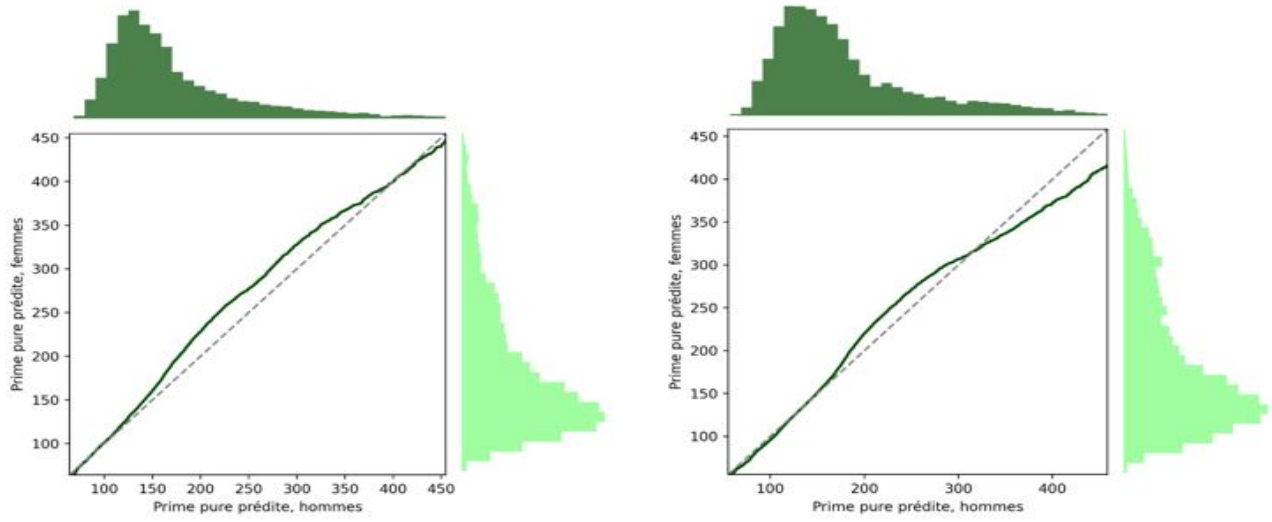

FIGURE 5.8 – Distributions conditionnelles de primes pures avant (à gauche) et après (à droite) orthogonalisation.

5.2.2 Mitigation par barycentre de distributions

Il s'agit de la méthode qui consiste à projeter chacune des primes pures conditionnelles à D dans un espace probabilisé ayant comme une mesure indépendante, mais barycentrique aux conditionnelles (Voir section 3.3). Elle a l'avantage de la simplicité et de sa réalisation peu coûteuse en temps de calcul. Les fonctions de répartition et de répartion inverse sont estimées à partir des prédictions de l'échantillon d'entrainement. La transformation se fait uniquement sur l'échantillon de test, et le modèle de départ est le « best-estimate », c'est-à dire celui utilisant explicitement la variable sensible.

Les résultats obtenus montrent une parité démographique au sens strict visuellement perceptible et attestée par les p-valeurs des tests de Kolmogorov-Smirnov (0,1923) et de Welch (0, 2851) significatives à 10%. Le graphique de correspondance montre une superpostion quasi-parfaite des distributions de primes pures prédites chez les hommes et chez les femmes. Comparativement à la méthode d'orthogonalisation, le niveau d'équité est donc meilleur ici. En effet, par rapport au modèle corrigé par la méthode précédente, la distance de Kolmogorov est diminuée de 80%, passant de 0,06 à 0,01. La différence de primes pures moyennes est réduite de moitié passant de $1,12\%$ à $0,56\%$.

En comparant point par point les primes pures initiales (obtenues au chapitre précédent dans le modèle ignorant 4.4.3) et transformées, on observe que cette méthode de mitigation entraine une diminution des primes pures chez les femmes et une légère augmentation de prime pure chez les hommes. Rappelons qu'initialement, les femmes paient en moyenne (10£ de) plus que les hommes (Voir gauche, figure 5.8). Pour atteindre la parité démographique, cette méthode de mitigation entraine une revue en augmentant légèrement les primes pures chez les hommes (en moyenne de 3€) contre une baisse chez les femmes (en moyenne de 10€). La partie droite du graphe ci-dessous l'illustre bien. Cette révision tarifaire [2] peut être comprise comme une forme de solidatité entre genres pour l'équité. Mais, quelles en seraient les implications sur le marché, n'affecterait-elle pas la souscription ? Répondre à ces questions nécessiterait une étude sur les dynamiques de marché, avec au préalable une méthode pour le passage à la prime commerciale.

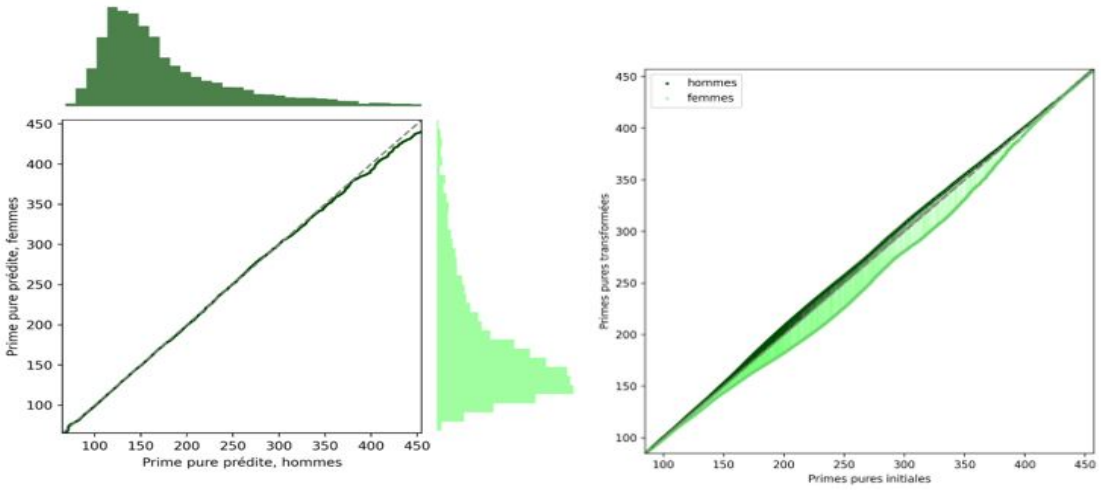

FIGURE 5.9 – Distributions conditionnelles de primes pures après mitigation (gauche), et comparaison entre prime pure initiale et prime pure équitable chez hommes et femmes (droite)

2. En faisant abstraction du passage à la prime commerciale

5.3 Discussion et recommandations

Les enseignements suivants peuvent être tirés des expérimentations précédentes :

— **La suppression de la variable sensible n'élimine pas tous les biais :**

Bien que la méthode « unaware » (ignorant la variable sensible) ait permis de réduire les disparités entre hommes et femmes, elle n'a pas suffi à atteindre une véritable parité démographique. Cela met en lumière l'existence de proxys indirects dans les données, tels que `ageph`, `bm`, ou `power`, qui peuvent refléter partiellement l'information liée au genre. Ces proxys, tout en étant légitimes dans la modélisation, contribuent à maintenir des biais structurels dans les prédictions.

— **L'orthogonalisation peut améliorer simultanément l'équité et les performances dans certains contextes :**

Les résultats montrent que l'orthogonalisation des variables explicatives vis-à-vis de la variable sensible peut réduire efficacement les disparités tout en augmentant, dans certains cas, les performances des modèles. Par exemple, pour les modèles de fréquence, une orthogonalisation complète ($\alpha = 1$) a permis d'améliorer la parité démographique tout en limitant l'impact sur la performance. Cependant, pour les modèles de coût, une orthogonalisation excessive a entraîné une dégradation de l'équité au-delà d'un certain seuil ($\alpha > 0.77$). Cela souligne la nécessité d'un équilibre subtil dans le choix de ce paramètre.

— **La méthode post-traitement par barycentre de distributions offre une correction puissante et interprétable :**

En partant des prédictions d'un modèle « best-estimate » (incluant la variable sensible), la méthode de transport optimal a permis d'atteindre une parité démographique quasi-parfaite, validée par les tests statistiques (Kolmogorov-Smirnov et Welch). Les résultats montrent une réduction significative des disparités moyennes entre genres (de 6.35% à 0.56%). Cette méthode, bien que très efficace en termes d'équité, nécessite une évaluation approfondie des impacts sur les primes commerciales et la dynamique de souscription sur le marché.

Les méthodes de mitigation explorées ici ne considèrent comme critère d'équité que le principe d'indépendance. La recherche d'autres principes nécessiterait l'utilisation de méthodes pendant la modélisation qui sont difficiles à calibrer et ne garantissent pas toujours un niveau de performance minimal. De plus, les métriques d'équité utilisées correspondent à des principes d'équité pas facilement compréhensibles ni recherchées à l'heure actuelle par la réglementation.

Conclusion

Ce mémoire s'est fixé pour objectif d'explorer les enjeux d'équité dans les modèles prédictifs appliqués à l'assurance, avec une attention particulière sur les discriminations involontaires liées à des caractéristiques sensibles comme le genre. À travers une analyse rigoureuse et des expérimentations pratiques, plusieurs contributions majeures ont été apportées à cette problématique complexe.

Dans un premier temps, les bases conceptuelles de l'équité ont été définies, en mettant en lumière la diversité des approches (individuelles et de groupe) et leurs implications dans le cadre des modèles actuariels. Ces définitions ont permis de cadrer l'analyse des biais potentiels, notamment ceux introduits par des proxys non-intentionnels des variables sensibles. Un cadre méthodologique a été proposé pour évaluer l'équité des modèles, intégrant des métriques comme la distance de Kolmogorov-Smirnov et la différence de parité statistique (`SPD`), qui s'avèrent pertinentes pour analyser les disparités dans les prédictions.

Ensuite, les expérimentations sur des modèles appliqués à l'assurance automobile ont permis d'illustrer concrètement les défis de l'équité. Les résultats obtenus montrent que la suppression des variables sensibles, bien que souvent perçue comme une solution intuitive, ne suffit pas à éliminer les biais dans les prédictions. En effet, des proxys, comme l'âge du véhicule ou le bonus-malus, peuvent continuer à capter des informations liées à la variable sensible. Par ailleurs, l'étude comparative des modèles avec et sans la variable sensible a révélé les tensions inhérentes entre équité et performance prédictive.

Les méthodes de mitigation explorées dans ce mémoire, notamment l'orthogonalisation des variables explicatives et l'ajustement post-modélisation par transport optimal, offrent des pistes concrètes pour réduire les disparités tout en préservant la pertinence des modèles. L'orthogonalisation s'est révélée efficace pour améliorer l'équité dans une certaine mesure, bien qu'un excès d'application puisse perturber la structure des données et entraîner des effets contre-productifs. Le transport optimal, quant à lui, a démontré une capacité supérieure à atteindre la parité démographique, au prix toutefois d'une légère perte de performance.

Les enseignements tirés de ce mémoire offrent des perspectives intéressantes pour les futurs travaux. Une piste prometteuse serait de combiner les approches pré-, in- et post-modélisation afin de maximiser simultanément équité et efficacité prédictive. De plus, l'intégration de notions causales dans l'évaluation des proxys pourrait renforcer la robustesse des analyses et permettre une meilleure interprétation des résultats. Enfin, l'exploration des impacts de ces ajustements sur les dynamiques de marché et la satisfaction des assurés reste un domaine à investiguer pour évaluer les implications économiques et sociétales des méthodologies proposées.

En conclusion, ce mémoire peut contribuer à une meilleure compréhension de l'équité dans le domaine assurantiel, tout en proposant des solutions pratiques pour répondre aux exigences éthiques et réglementaires. Ces travaux s'inscrivent dans une dynamique plus large visant à promouvoir une utilisation responsable et transparente de l'intelligence artificielle dans un secteur en pleine mutation.

Références Bibliographiques

[1] Alekh Agarwal, Miroslav Dudík et Zhiwei Steven Wu. "Fair regression : quantitative definitions and reduction-based algorithms." In : *In ICML, volume 97 of Proceedings of Machine Learning Research, 120–129. PMLR* (2019).

[2] Alekh Agarwal et al. "A Reductions Approach to Fair Classification". In : *Proceedings of the 35th International Conference on Machine Learning, PMLR 80 :60-69* (2018).

[3] Richard Berk et al. "A convex Framework for fair regressions". In : *Arxiv 1706.02409* (2017).

[4] Alessandro C. et al. "A clarification in the fairness metrics landscape". In : *Scientific reports, Sci 12, 4209* (2022).

[5] Arthur Charpentier. *Insurance, Biases, Discriminations and Fairness*. Springer Actuarial, 2024.

[6] Arthur Charpentier, François Hu et Philipp Ratz. "Mitigating Discrimination in Insurance with Wasserstein Barycenters". In : *DOI :10.48550/arXiv.2306.12912* (2023).

[7] European commission. *EU rules on gender-neutral pricing in insurance*. Consulté le 07 Juillet 2024, `https://ec.europa.eu/commission/presscorner/detail/en/MEMO_12_1012`. 2012.

[8] European commission. *Sex Discrimination in Insurance Contracts : Statement by European Commission Vice-President Viviane Reding, the EU's Justice Commissioner, on the European Court of Justice's ruling in the Test-Achats case*. consulté le 07 Juillet 2024, `https://ec.europa.eu/commission/presscorner/detail/en/MEMO_11_123`. 2011.

[9] European commission et Groupe d'experts. *Lignes Directrices en matière d'équité pour une IA digne de confiance*. 2018.

[10] Olivier Côté, Marie-Pier Côté et Arthur Charpentier. "A Fair price to pay : exploiting causal graphs for fairness in insurance". In : *Available as SSRN : https://ssrn.com/abstract=4709243 or http://dx.doi.org/10.2139/ssrn.4709243* (2024).

[11] Freddy Delbaen et Chitro Majumdar. "APPROXIMATION WITH INDEPENDENT VARIABLES". In : *Frontiers of Mathematical Finance 2(2) : 141-149, doi : 10.3934/fmf.2023011* (2023).

[12] Christophe Dutang et Arthur Charpentier. "CASdatasets : Insurance datasets, R package version 1.2-0". In : *DOI 10.57745/P0KHAG* (2024).

[13] Cynthia Dwork et al. "Fairness Through Awareness". In : *Arxiv 1104.3913* (2011).

[14] Louis Houde. "Tests du Khi-deux". In : *Département de Mathématiques et d'informatique, UQTR* (2014).

[15] Michael W. Kearney. "Cramer's V". In : *Sage Encyclopedia of Communication Research Methods* (2017).

[16] Issa Kohler-Hausmann. "Discrimination". In : *Oxford Bibliographies Online, https: www. oxfordbibliographies. com/ display/ document/ obo-9780199756384/ obo-9780199756384-0013. xml , Consulté le 20 Juillet 2024* (2011).

[17] William H. Kruskal et W. Allen Wallis. "Use of ranks in one-criterion variance analysis". In : *Journal of the American statistical association* (1952).

[18] Matt Kusner et al. "Counterfactual Fairness". In : *Arxiv 1703.06856v1* (2018).

[19] Xavier Landes. "How Fair Is Actuarial Fairness ?" In : *J Bus Ethics 128, 519–533* (2015).

[20] Mathias Lindholm et al. "Sensitivity-based measures of discrimination in insurance pricing". In : *SSRN : https: // ssrn. com/ abstract=4897265* (2024).

[21] Mathias Lindholm et al. "What is fair ? Proxy discrimination vs demographic disparities in insurance pricing". In : *Scandinavian Actuarial Journal, DOI : 10.1080/03461238.2024.2364741* (2024).

[22] *Manuel de droit européen en matière de non-discrimination, édition 2018.*

[23] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In : *Arxiv 1908.09635v3* (2022).

[24] Mulah Moriah. "Mesure et mitigation des biais : vers une tarification non-vie réellement équitable". In : *Institut des Actuaires* (2022).

[25] Ferignac P. "Test de Kolmogorov-Smirnov sur la validité d'une fonction de distribution". In : *Revue de la statistique appliquée, tome 10, no 4* (1962).

[26] Ricco Rakotomalala. "Comparaison de populations, Tests non paramétriques (P11-P23)." Université Lumière Lyon 2.

[27] John Rawls. *Théorie de la Justice*. Bookey - Accès : `https://www.bookey.app/fr/book/theorie-de-la-justice`. Résumé disponible sur Bookey. 2024.

[28] Boris Ruff et Marcin Detyniecki. *Towards the Right Kind of Fairness in AI*. Rapp. tech. 2021.

[29] Kullback S. et Leibler R. “On information and sufficiency”. In : *The Annals of Mathematical Statistics, Vol. 22, No. 1* (1951).

[30] Marguerite Saucé. “AI and ethics in insurance : a new solution to mitigate proxy discrimination in risk modeling”. In : *Institut des actuaires* (2023).

[31] Steven Sawyer. “Analysis of Variance : The Fundamental Concepts”. In : *The Journal of manual manipulative therapy* (2009).

[32] Gero Szepannek et Karsten Lübke. “Facing the Challenges of Developing Fair Risk Scoring Models”. In : *frontiers in Artifical intelligence, doi : 10.3389/frai.2 021.681915, p 1 - 4* (2021).

[33] Brian Hu Zhang, Blake Lemoine et Margaret Mitchell. “Mitigatin Unwanted Biases with Adversarial Learning”. In : *AAAI/ACM Conference on AI, Ethics, and Society (AIES ’18), February 2–3, 2018, New Orleans, LA, USA. ACM, New York, NY, USA, 6 pages.* `https://doi.org/10.1145/3278721.3278779` (2018).

Annexes

Principes d'équité

Arbre de décision

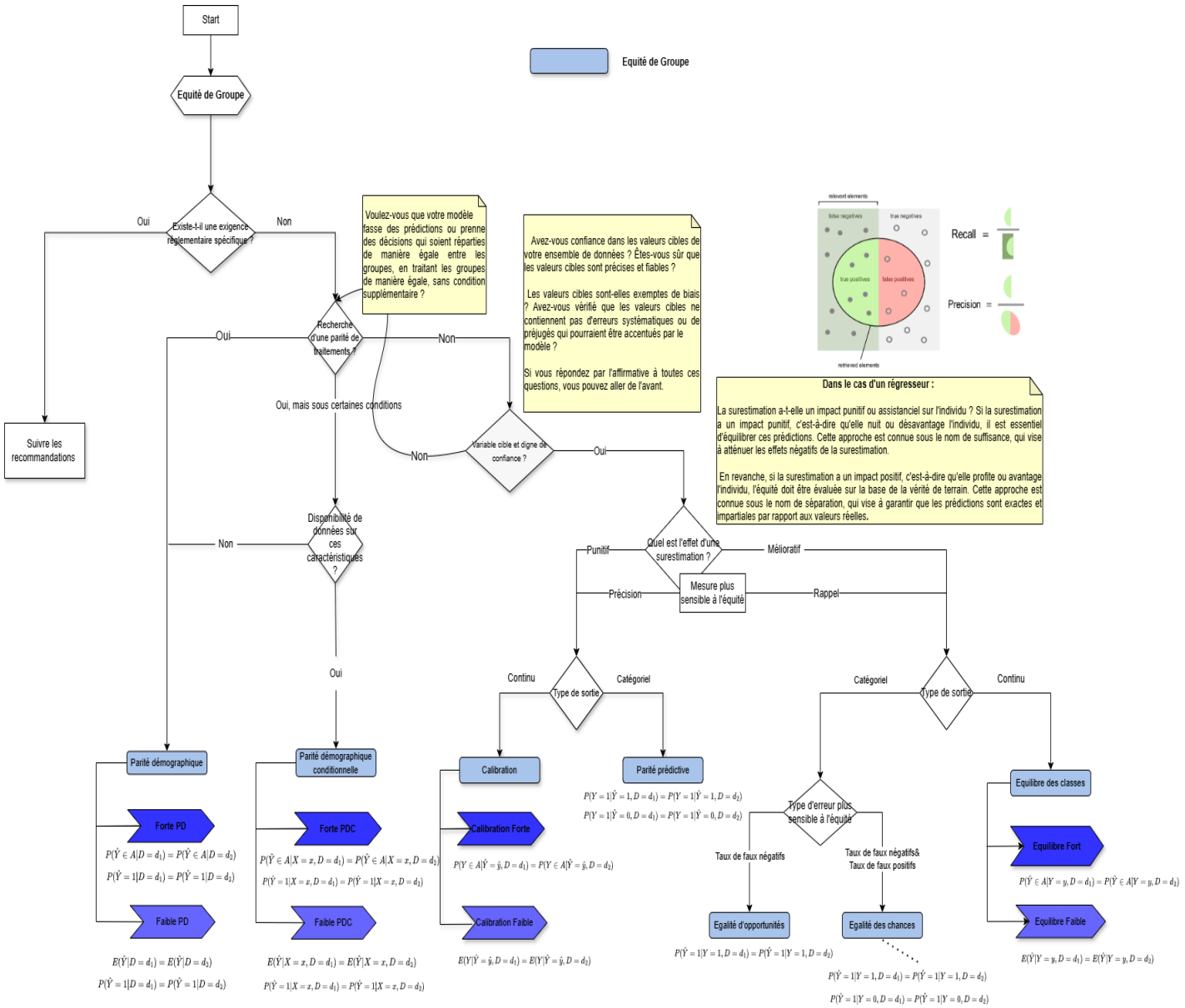

FIGURE 5.10 – Décison sur les principes d'équité de groupe

Evaluation des métriques sur données simulées

Comparaison des modèles dans un cadre statique

— **Illustrations de la vérification de certains critères d'équité de groupe - Indépendance**

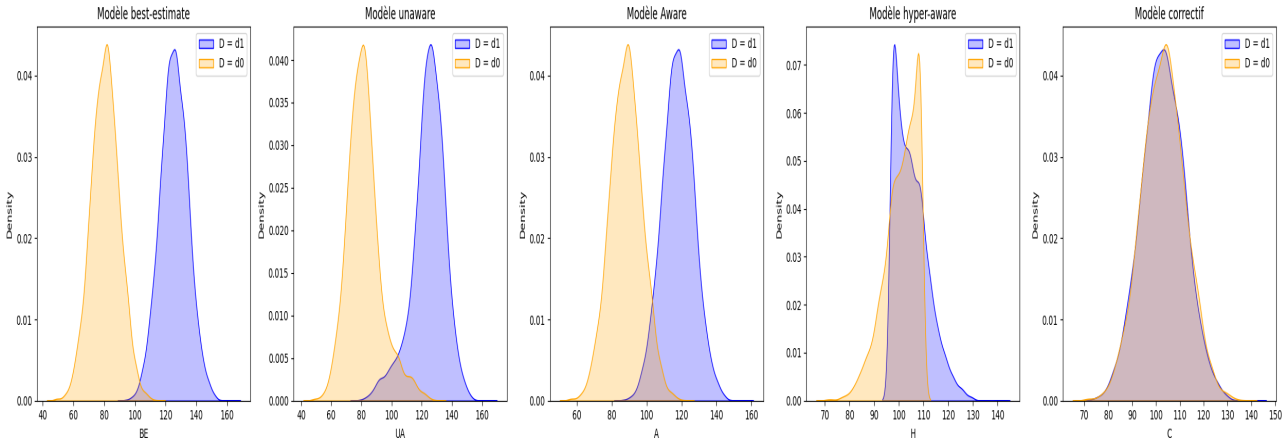

FIGURE 5.11 – Distributions des prédictions de différents modèles conditionnellement à D

De gauche à droite, on a les distributions de prédictions obtenues pour les modèles best-estimate, unaware, aware, hyper-aware et correectif. Les distibutions dans les groupes de D se confondent pour ce dernier modèle, ce qui illustre qu'il satisfait le critère d'indépendance.

— **Illustrations de la vérification de certains critères d'équité de groupe - Séparation**

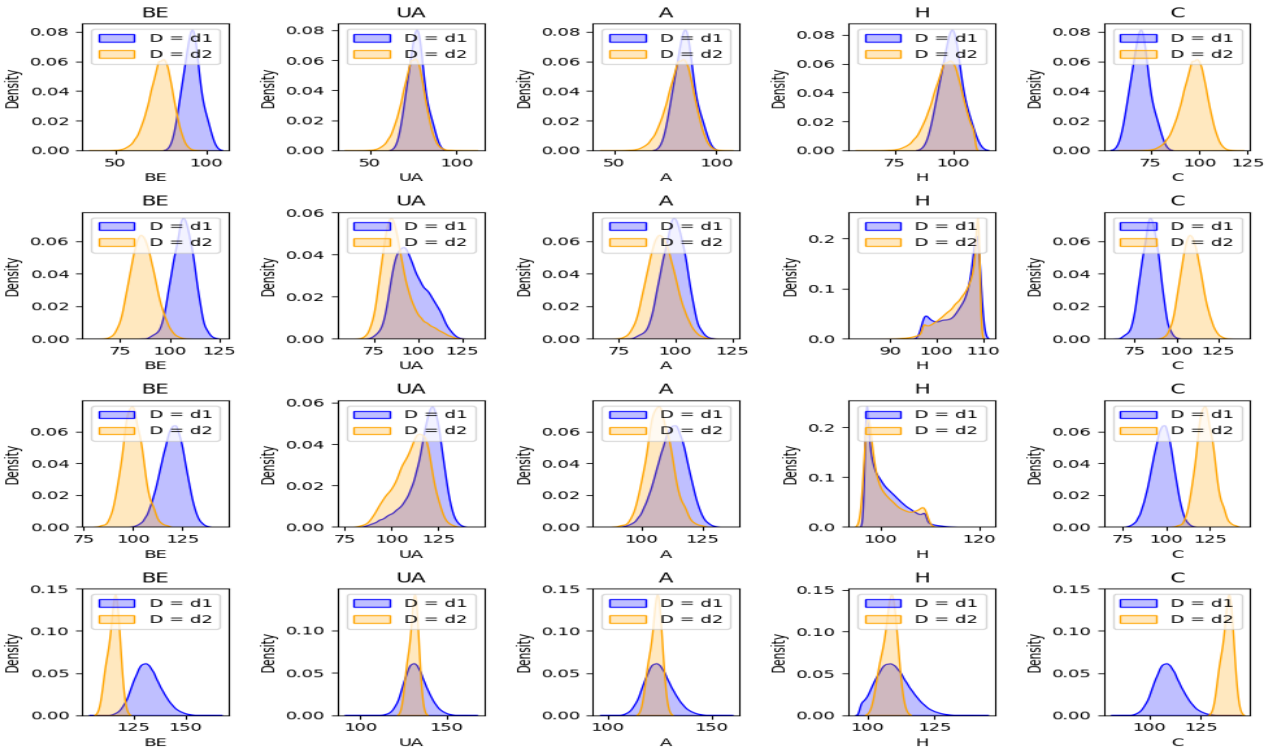

FIGURE 5.12 – Distributions des prédictions de différents modèles dans les catégories de D conditionnellement à Y

De gauche à droite, on a les cinq modèles précédents dans le même ordre. Les lignes représentent les intervalles de Y définis par les quartiles de sa distribution empirique. Le modèle Hyper-averti est le plus équitable au sens de la séparation

Evaluation de la sensibilité des métriques dans un cadre dynamique

Paramètre	Description	Valeur	Remarques
k	Coefficient de D dans X_1	-	Paramètre influent sur X_1 et Y
μ	Moyenne de $\mathcal{N}(\mu, \sigma^2)$	0	Valeur de la moyenne du bruit dans X_1
σ^2	Variance de $\mathcal{N}(\mu, \sigma^2)$	1	Variance du bruit dans X_1
λ	Paramètre de l'exponentielle pour Z	0.5	Régit la distribution de $Z \sim E(\lambda)$
α_1	Coefficient de X_1 dans Y	1	Coefficient linéaire sur X_1 dans Y
α_2	Coefficient exponentiel dans Y	Ajusté	Paramètre influent sur l'effet de X_1 dans Y
α_3	Coefficient de Z dans Y	0.2	Coefficient linéaire sur Z dans Y
α_4	Coefficient de D dans Y	0.5	Paramètre influent sur D dans Y
ϵ	Bruit gaussien dans Y	$\mathcal{N}(0, \epsilon^2)$	Bruit ajouté dans le modèle de Y

TABLE 5.1 – Résumé des valeurs des paramètres

Description de la base de données réelles

Analyse multidimensionnelle des variables

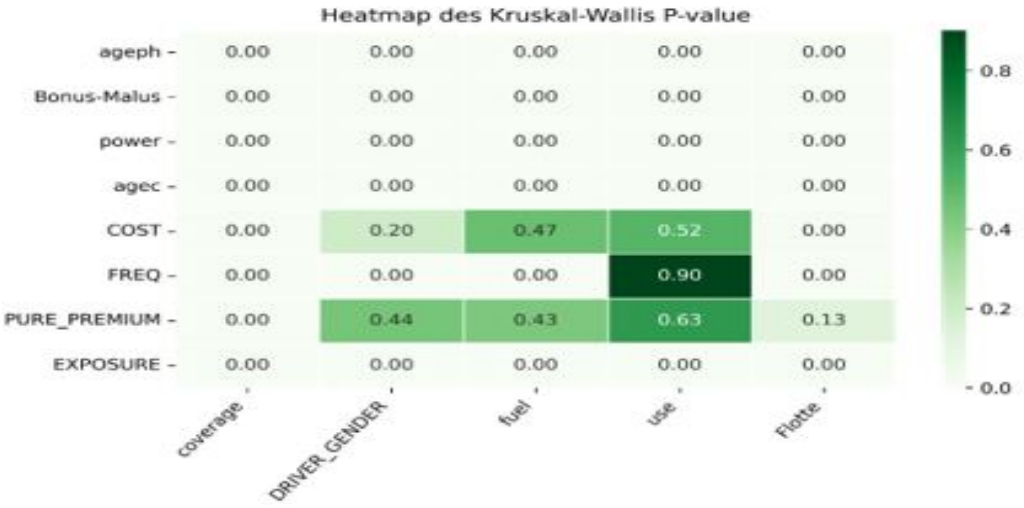

FIGURE 5.13 – P-valeurs du test de KrusKal Wallis

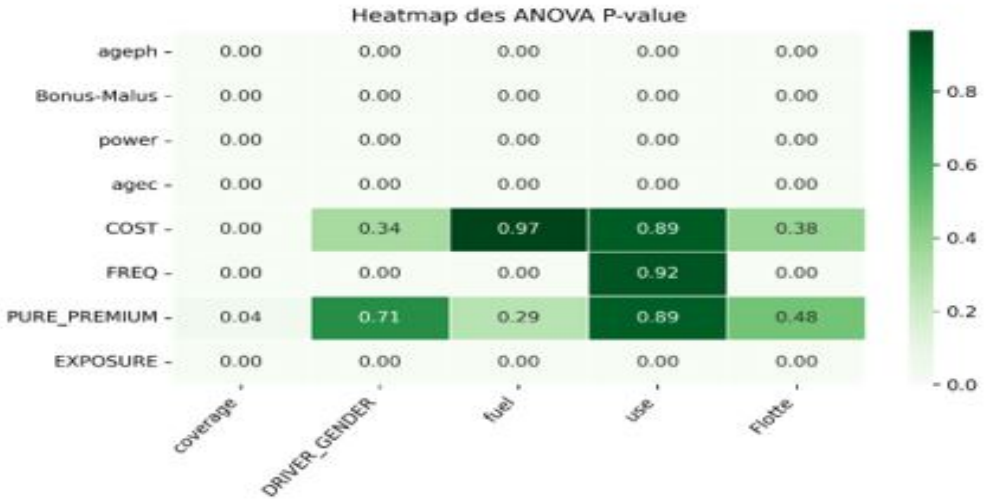

FIGURE 5.14 – P-valeurs du test d'ANOVA

Examen des méthodes de mitigation

L'orthogonalisation des variables explicatives

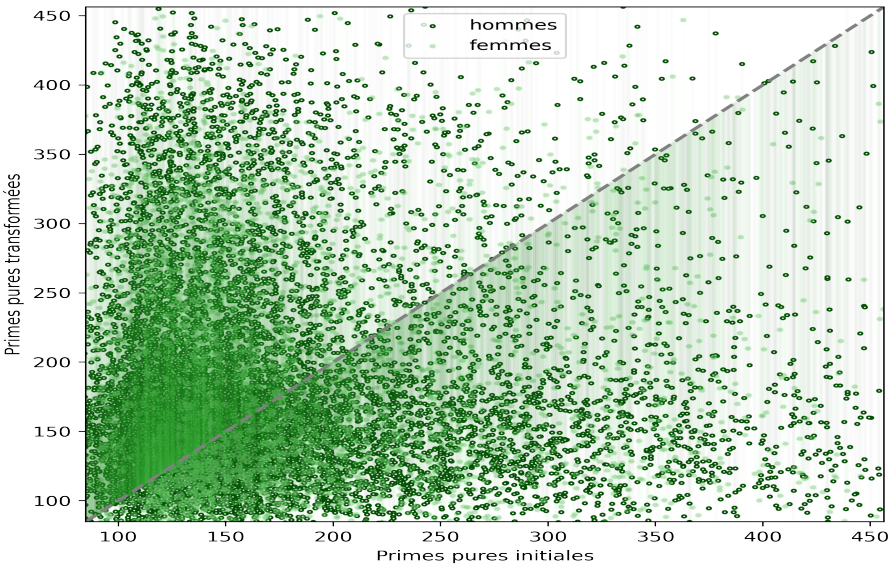

FIGURE 5.15 – Primes pures avant vs après orthogonalisation

Les changements sont importants et il n'y a pas de tendance générale qui se dégage