UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÈLES ÉQUITABLES DE TARIFICATION EN ASSURANCE AUTOMOBILE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

CHRISTINA OUEINI

NOVEMBRE 2023

REMERCIEMENTS

Je tiens particulièrement à remercier mon directeur de recherche, Mathieu Pigeon, sans qui ce travail n'aurait pas pu être possible. Son expertise, ses conseils et son support ont été d'une aide précieuse et m'ont permis de développer mes connaissances et mon esprit analytique.

Je souhaite également remercier la Compagnie d'assurance générale Co-Operators de m'avoir permis d'accéder à leurs données. Un grand merci à Étienne Larrivée-Hardy, directeur de l'équipe de recherche et innovations analytiques à Co-Operators, qui a partagé ses idées et sa vision afin de combiner la théorie avec la pratique. Je remercie aussi la Chaire Co-Operators en analyse des risques actuariels à l'UQAM d'avoir rendu cette collaboration possible et pour l'aide financière offerte pour le soutien de la poursuite de mes études.

Je remercie aussi les évaluateurs qui, grâce à leur connaissance du domaine, ont pu rajouter de la valeur à ce mémoire de recherche.

Finalement, à mes parents, qui m'ont motivé et encouragé tout au long de mon parcours académique. J'en suis profondément reconnaissante.

Alors tu comprendras la justice, l'équité, la droiture,
toutes les routes qui mènent au bien.
- Proverbes 2 :9

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

L'apprentissage machine est largement répandu dans l'industrie de l'assurance. Un exemple de son utilisation dans ce milieu est pour le calcul de la prime des assurés. La prime doit être représentative du niveau de risque de ce dernier et le quantifier de la manière la plus juste possible. La question d'équité dans les modèles d'apprentissage automatiques a été soulevée à la suite de certains incidents discriminatoires qui ont surgi dans le passé. En effet, durant le processus d'entrainement du modèle, certaines variables sensibles, comme le sexe, l'âge ou l'état civil, sont utilisées et celles-ci peuvent mener à des résultats discriminatoires envers les groupes minoritaires. Les assureurs sont les seuls organismes qui ont encore le droit de posséder et d'utiliser de telles informations, mais leur pratique a récemment soulevé beaucoup de questionnement éthique (Commission *et al.*, 1999). Les régulateurs encouragent l'utilisation de ces variables de manière non discriminatoire et le présent travail de recherche propose un modèle de tarification automobile qui mitige la divergence entre les résultats des sous-groupes appartenant à un certain attribut sensible. Le modèle de tarification en question incorpore une contrainte d'équité qui permet de contrôler l'effet des variables sensibles utilisées dans l'entrainement. La divergence entre le modèle sans contrainte et avec contrainte est calculée grâce à une métrique de quantification de l'équité. Les résultats sont concluants et démontrent qu'avec l'ajout d'une contrainte d'équité, il est possible d'obtenir des primes équitables qui respectent la définition établie dans le contexte de ce travail.

Mots clés : Discrimination, Apprentissage machine, Tarification, Contrainte d'équité, Assurance automobile.

ABSTRACT

Machine learning is widespread in the insurance industry. An example of its use in this environment is for the calculation of the premium of the insured. The premium must be representative of the latter's level of risk and quantify it as accurately as possible. The question of fairness in machine learning models has been raised following some discriminatory incidents that have arisen in the past. Indeed, during the model training process, certain sensitive variables, such as gender, age, or marital status, are used and these can lead to discriminatory results against minority groups. Insurers are the only organizations that still have the right to possess and use such information, but their practice has recently raised a lot of ethical questions (Commission *et al.*, 1999). Regulators encourage the use of these variables in a non-discriminatory way and this research work proposes an automobile pricing model that mitigates the discrepancy between the results of subgroups belonging to a certain sensitive attribute. The pricing model in question incorporates an equity constraint that controls the effect of sensitive variables used in training. The divergence between the unconstrained and constrained model is calculated using an equity quantification metric. The results are conclusive and demonstrate that with the addition of an equity constraint, it is possible to obtain fair premiums that respect the definition established in the context of this work.

Key words : Discrimination, Machine learning, Pricing, Equity constraint, Auto insurance.

INTRODUCTION

Les décisions prises par l'entremise des algorithmes d'apprentissage statistique deviennent de plus en plus fréquentes dans les industries. Par exemple, dans le domaine de l'assurance, ces algorithmes sont utilisés pour le calcul des primes des différentes couvertures. Les prédictions faites par ces algorithmes peuvent, cependant, apprendre à discriminer négativement certaines personnes ou certains groupes de personnes. En effet, l'utilisation de données historiques, qui contiennent parfois des biais, pour l'entrainement amène ces algorithmes à apprendre la discrimination et ceci est reflété dans les résultats. Parfois, il y a même un manque de représentation dans les données pour certaines minorités et le modèle n'est pas en mesure de représenter ce groupe de manière adéquate et juste dans les prédictions.

En assurance, la discrimination est un terme neutre puisqu'il fait allusion à la segmentation des assurés. Cette segmentation nous permet de classifier les assurés selon leur risque et, ainsi, leur attribuer une prime qui reflète leur niveau de risque respectif. Ce travail est effectué sur des bases de données historiques qui peuvent contenir un certain type de biais, surtout envers les minorités. Ceci fait en sorte que cette prime pourrait être discriminatoire envers les individus appartenant à cette minorité. Les types de biais qui peuvent être présents dans les données sont, par exemple, les biais historiques ou bien, il est même possible d'avoir un manque de données pour représenter adéquatement les minorités.

Le présent travail de recherche vise à explorer une façon de contrôler le biais que l'on pourrait retrouver dans les bases de données utilisées pour l'apprentissage d'un algorithme de tarification en assurance automobile. L'objectif ultime à atteindre

est de faire en sorte que la prime affectée à un assuré reflète complètement le risque de ce dernier et n'est pas influencée par une variable sensible quelconque telle que l'âge ou le sexe.

Pour se faire, quatre modèles de tarification ont été entrainés et analysés. On commence par modéliser la prime pure par l'entremise d'un modèle linéaire généralisé (MLG) appartenant à la famille Tweedie qu'on compare avec un modèle Boosting Tweedie en ajoutant une contrainte d'équité qui vise à contrôler l'impact de la variable sensible sur les prédictions du modèle. Ensuite, la modélisation de la fréquence et de la sévérité sont réalisées séparément et la contrainte d'équité est rajoutée dans la modélisation de la fréquence. Ainsi, un MLG pour la fréquence et un MLG pour la sévérité sont mis en oeuvre, de même qu'un algorithme de Boosting Bernoulli pour la fréquence uniquement.

Au chapitre 1 des notions théoriques d'équité sont présentées pour aider à la compréhension de la problématique et pour établir les connaissances nécessaires pour la suite de ce travail de recherche. La discrimination dans les algorithmes d'apprentissage machine et la manière dont celle-ci affecte l'industrie de l'assurance sont abordées.

Le chapitre 2 revoit la littérature à ce sujet et introduit plusieurs méthodes d'intervention pour la mitigation de la discrimination dans les algorithmes d'apprentissage machine. L'équité en assurance est un sujet de recherche assez récent et la documentation à ce sujet est encore peu élaborée. Ceci explique l'utilisation de certains preprints pour la revue de littérature de ce travail. Dans ce chapitre, les concepts de discrimination en actuariat sont expliqués. Ceci est suivi par la présentation de plusieurs métriques d'équité qui peuvent être incorporées dans les algorithmes de classification ou de régression. Celle qui est retenue est présentée dans ce chapitre et sa justification est définie. Finalement, on termine avec les

différents moments dans le processus de modélisation où il est possible de mitiger la discrimination.

Le chapitre 3 aborde la théorie des modèles linéaires généralisés et des algorithmes de Boosting. La théorie des modèles construits précisément dans le cadre de cette recherche est également présentée. De plus, l'intégration de la contrainte d'équité choisie dans l'algorithme de Boosting est expliquée plus en détail.

Le chapitre 4 présente les résultats numériques obtenus. Une description explicite de la base de données utilisée pour l'entrainement des modèles est faite. La technique de sélection de variables est expliquée et la comparaison de la performance des modèles est effectuée. Le niveau de discrimination par rapport aux variables sensibles choisies est quantifié à l'aide des métriques abordées dans les chapitres précédents et, finalement, le calcul des primes pures selon chaque modèle sont illustrées pour quelques profils.

On conclut ce mémoire en suggérant l'exploration d'une différente approche pour améliorer les lacunes de ce travail, tout en gardant en tête le but principal d'obtenir des primes équitables pour tous.

CHAPITRE I

ÉQUITÉ EN ASSURANCE

1.1 Problématique

La discrimination systémique est une préoccupation dans notre société (Gosselin, 2020) et il est critique de s'assurer que ce traitement inéquitable n'est plus perpétué, surtout à la suite de l'arrivée de l'apprentissage machine dans plusieurs industries. Les individus appartenant à un groupe minoritaire sont souvent représentés de manière inadéquate et biaisée (Dickey, 2020). Ceci peut être reflété dans les données récoltées par les entreprises qui utilisent ces mêmes données dans la construction d'algorithmes prédictifs qui assistent dans certaines prises de décisions. L'utilisation de modèles statistiques est de plus en plus répandue dans plusieurs domaines comme la sélection d'emploi, la prédiction de récidive ou la détection de fraude. Il est crucial de faire en sorte que les méthodes utilisées dans ces modèles soient justes et fiables.

Le domaine de l'actuariat est également une industrie qui base souvent ses décisions sur des modèles statistiques. Ce qui distingue l'assurance commerciale des autres domaines est le fait qu'elle est notamment fondée sur la discrimination des risques de chaque assuré et l'atteinte des objectifs de l'entreprise est basée sur ce principe. La compétition entre les assureurs oblige les compagnies à segmenter pour offrir des produits abordables aux assurés. Par crainte d'antisélection, les

assureurs tentent d'avoir la meilleure segmentation pour le regroupement de leurs assurés. Le terme discrimination a souvent une connotation négative mais, en assurance, c'est plutôt un terme neutre qui désigne la distinction entre les risques de chaque individu, ce qui fait partie des normes actuarielles fondamentales. Néanmoins, cette pratique peut venir à l'encontre des droits de l'homme qui interdisent la discrimination sur les classes protégées suivantes : le handicap, l'âge, le sexe et l'état civil (Ontario Human Rights Commission, 2005). Une classe protégée, ou un attribut sensible, est une variable qui, historiquement, a été utilisée dans le but d'administrer un traitement intentionnellement défavorable aux individus qui y appartenaient (Haeri et Zweig, 2020). Une classe protégée est donc un groupe de personnes qui possède une caractéristique commune et qui est protégé contre la discrimination par la loi. L'assurance est la seule industrie où la discrimination est encore permise sous la condition que celle-ci soit fondée sur le risque (Commission *et al.*, 1999). Autrement dit, si une distinction sur une classe protégée est utilisée dans la segmentation pour le calcul de la prime d'un assuré, elle doit être représentative et proportionnelle au risque de cet assuré. Si ce n'est pas le cas, il risque d'y avoir un traitement défavorable envers un groupe minoritaire et ceci n'est pas permis par la loi.

Ces discussions sont de plus en plus présentes au sein de l'industrie afin de s'assurer que tous ont accès à une prime équitable. Les travaux de recherche à ce sujet prennent de l'ampleur afin de trouver des mesures de tarification alternatives tout en respectant les différentes définitions d'équité que l'industrie cherche à honorer. L'implication des régulateurs est d'une haute importance afin que toutes les compagnies d'assurance aient la même base en matière de connaissances et de réglementation afin d'éviter l'antisélection. Ce sont des mesures qui deviennent de plus en plus concrètes dans le monde de l'assurance et la mise en oeuvre de modèle de tarification équitable se voit devenir la norme de plus en plus.

La recherche effectuée dans le cadre de ce mémoire aborde ce sujet dans le domaine de la science actuarielle et explore une possible avenue pour contrôler l'effet discriminatoire des attributs sensibles dans les algorithmes de tarification en assurance automobile.

1.2 Discrimination dans les algorithmes d'apprentissage machine

L'utilisation de données historiques dans les algorithmes d'apprentissage machine est remise en question puisque que ces dernières ne sont pas toujours justes et complètes à l'égard des groupes minoritaires. En utilisant ces données pour construire des modèles, les algorithmes sont en mesure d'apprendre à discriminer négativement, même si c'est de manière involontaire, les individus qui font partie d'un groupe minoritaire.

La discrimination dans les décisions automatisées peut survenir pour deux raisons principales. La première source de biais est le manque de données pour représenter les minorités (Dickey, 2020). En effet, si un algorithme d'apprentissage machine est entrainé sur peu de données, il ne sera pas en mesure de bien capter la relation statistique présente entre les variables explicatives et la variable réponse et les prédictions ne seront donc pas précises. De même, si dans la base de données il y en a peu pour représenter une sous-classe d'un attribut sensible, le modèle ne sera pas juste pour les individus appartenant à cette sous-classe. Il y aura une disparité dans la performance du modèle qui pourra affecter, négativement ou positivement, les sous-classes des attributs protégés.

La deuxième source de biais est celle qui peut se trouver dans les bases de données utilisées. Les données qui sont utilisées pour entrainer les modèles reflètent, parfois, des représentations erronées de certains groupes (Zliobaite, 2015) ce qui peut, en conséquence, donner des prédictions erronées également. Les décisions prises par

l'intelligence artificielle ne sont pas toujours objectives et la discrimination peut être captée à travers les relations sous-jacentes entre les variables explicatives. Le processus d'apprentissage peut même aggraver ce phénomème dans le cas de sur-ajustement du modèle (Wang *et al.*, 2020). La compréhension des données et des modèles est donc essentielle pour être en mesure d'ajuster un algorithme précis et juste pour tous.

1.3 Type de discrimination

L'inéquité dans les modèles statistiques peut être captée de deux manières distinctes : soit en capturant le lien direct entre la variable réponse et l'attribut sensible, soit à travers une relation indirecte entre la variable réponse et une variable corrélée à l'attribut sensible. Il est aussi important de noter que, dans le cadre de ce travail de recherche, on parle notamment de discrimination négative.

1.3.1 Discrimination directe

La discrimination directe survient lorsqu'une personne, ou un groupe de personnes, est traitée défavorablement en raison de son appartenance à un attribut protégé. Par exemple, l'ethnicité, le sexe, l'âge ou bien l'orientation sexuelle sont des variables sur lesquelles la discrimination directe peut survenir. En effet, si l'une de ces variables se trouve dans un modèle d'apprentissage statistique, les résultats issus de ce modèle peuvent être inéquitables envers ces groupes minoritaires s'ils sont mal représentés dans la base de données. Ceci est la forme de discrimination la plus apparente. C'est également celle qui est la plus facile à contrôler. En effet, pour éliminer la discrimination directe d'un modèle, il suffit d'ôter la variable en question durant le processus d'entrainement.

1.3.2 Discrimination indirecte

La discrimination indirecte, quant à elle, survient lors de l'utilisation de variables proxy dans le modèle. Une variable proxy est une variable considérablement corrélée à l'attribut sensible et qui pourrait même être utilisée pour le remplacer. Par ailleurs, même si les attributs sensibles ne font pas partie des variables utilisées pour l'entrainement du modèle, la discrimination peut tout de même être apprises par le biais des variables qui sont corrélées aux caractéristiques protégées. Par exemple, on pourrait avoir le code postal qui est souvent un proxy pour l'ethnicité. Souvent, l'ethnicité est l'une des variables interdites dans plusieurs industries, mais si on possède le code postal d'un individu, ce qui est souvent le cas en assurance automobile, il est possible de discriminer sur l'ethnicité de cet individu malgré tout par le biais du territoire où il réside. Ce phénomène démontre que la suppression des variables considérées problématiques n'est pas suffisante pour arbitrer l'inéquité d'un modèle statistique. De plus, la suppression de tous les attributs sensibles ainsi que de leur proxy n'est pas une solution viable puisqu'on risquerait alors d'avoir très peu de variables avec lesquelles travailler. Par ailleurs, la segmentation en assurance est fondamentale pour modéliser le risque des assurés et si elle n'est pas accomplie de manière adéquate, il pourrait y avoir de l'antisélection, ce qui signifie qu'un assuré pourrait obtenir une prime inférieure ou supérieure à celle associée à son réel niveau de risque. Cela entrainerait une asymétrie de l'information entre compétiteurs et causerait le départ des assurés communément appelés les «bons» risques et une surreprésentation des assurés sous tarifés ou les «mauvais» risques. Des pertes financières pourraient en résulter. Le cas échéant, la tarification en assurance risque d'être moins précise et la solvabilité de la compagnie pourrait en souffrir. Puisque la suppression de toutes les variables proxy ne semble pas être une solution durable, une technique de contrôle de l'influence de cette même variable sur les prédictions sera présentée

ultérieurement dans le cadre de ce travail de recherche.

1.4 Diminution de la disparité entre groupe

Dans les algorithmes équitables, il existe principalement deux concepts d'équité. Premièrement, on retrouve l'équité individuelle et, deuxièmement, l'équité de groupe. En ce qui concerne l'équité individuelle, on cherche à obtenir un traitement équitable pour chaque individu. Des profils similaires devront donc produire des résultats similaires selon l'algorithme. Cette forme d'équité est plus difficile à mettre en oeuvre informatiquement puisqu'il se pourrait qu'il y ait plusieurs modalités à prendre en compte dans une certaine variable et à comparer avec chaque profil (Lohia *et al.*, 2019).

Deuxièmement, il est possible de créer des modèles équitables en respectant la définition de l'équité de groupe. Comme il a été mentionné précédemment, la segmentation permet de classer les assurés dans des groupes relatifs à leur niveau de risque. L'atténuation de la discrimination en assurance vise à faire en sorte que différents groupes soient traités équitablement. Pour se faire, on met en oeuvre une mesure de parité statistique qui devra être égale entre tous les sous-groupes appartenant à un attribut sensible quelconque (Lohia *et al.*, 2019). Ceci sera abordé plus en détails au chapitre suivant. De plus, l'écart entre les performances du modèle dans différents sous-groupes ne doit pas être significativement différent pour que l'algorithme soit considéré comme juste et équitable. La partition de ces sous-groupes est faite sur une classe protégée qui se trouve dans le modèle, comme par exemple, l'âge, le sexe ou l'ethnicité. En assurance, la forme d'équité qu'on cherche à satisfaire est bel et bien l'équité de groupe.

1.4.1 Discrimination en assurance

En utilisant l'information à leur disposition, les actuaires tentent de modéliser le risque en créant une tarification similaire entre les assurés appartenant à un même groupe. La concurrence entre les compagnies d'assurance pousse les assureurs à approfondir leur segmentation en allant chercher des variables de plus en plus spécifiques et ainsi à travailler avec des données davantage détaillées. La méthode de tarification en assurance est remise en question dans les années 60 alors que la lutte contre la discrimination prend de l'ampleur avec le début des efforts pour reconnaitre les droits civils des Afro-Américains aux États-Unis (Charpentier et Barry, 2022).

En 1934, la pratique du *redlining* a débuté aux États-Unis. Celle-ci consistait à délimiter les quartiers résidentiels selon leur niveau de désirabilité en leur assignant une couleur (vert, bleu, jaune ou rouge). Cette approche a été utilisée par l'administration fédérale du logement (FHA) afin de déterminer l'admissibilité à obtenir une assurance sur l'hypothèque d'une résidence (Chibanda, 2022). Cette approche a ensuite été critiquée pour cause de discrimination sur l'ethnicité puisque les quartiers identifiés comme étant les moins désirables étaient principalement ceux où des minorités résidaient. Le *redlining* est devenu illégal d'après le *Fair Housing Act* en 1968.

L'usage de la variable sexe en tarification a également apporté son lot de problèmes dans la population qui se bat pour l'équité. Le Montana est le premier état aux États-Unis, en 1985, à bannir son usage dans l'industrie de l'assurance à la suite des efforts du mouvement des groupes féministes dans la lutte pour une tarification unisexe (Reid et report, 1985). Toutes les couvertures d'assurance ont désormais une segmentation qui omet l'usage du sexe. Plusieurs états, tels que la Californie, Hawaii, le Massachusetts et le Michigan, ont suivi cette initiative en excluant cette

variable dans le calcul des primes en assurance automobile. L'Union européenne a également banni l'usage du sexe dans l'estimation des primes en 2012 et le calcul est dorénavant fait par l'entremise des variables qui sont directement en lien avec la conduite de l'assuré telles que la marque de la voiture et le kilométrage parcouru (Lichtenstein, 2022).

L'affaire de Bates v. Zurich est un cas qui démontre assez bien l'usage conflictuel des attributs sensibles. En Ontario, en 1983, Michael Bates a entamé une poursuite judiciaire contre la compagnie d'assurance Zurich prétendant que sa prime était discriminatoire sur l'âge, le sexe et l'état civil. Cependant, la commission ontarienne des droits de la personne stipule qu'il est légal de discriminer sur ces attributs tant que leur utilisation est nécessaire à la survie de la compagnie et qu'ils ne sont pas employés de manière défavorable envers l'assuré. Le jugement final a été établi selon deux conditions. Premièrement, il a été démontré que l'utilisation de ces attributs dans le cas de Bates n'est pas discriminatoire et, deuxièmement, il n'y avait pas de méthode alternative qui puisse remplacer l'usage de ces attributs dans la tarification à l'époque (Commission *et al.*, 1999). Depuis, de nouvelles techniques de tarification ont fait leur apparition et le début de l'utilisation de l'apprentissage machine en assurance a modernisé ces pratiques. La science actuarielle a énormément évolué dans les dernières années et la recherche en matière de tarification non discriminatoire ne cesse d'évoluer.

L'apprentissage machine fait son entrée aux États-Unis dans le monde de l'assurance dans les années 1990 avec la mise en oeuvre des modèles linéaires généralisés (GLM) comme méthode de modélisation prédictive en utilisant des données historiques (Werner et Guven, 2007). Son utilisation a fait une différence dans la précision des résultats, qui ont généré des profits supplémentaires pour les compagnies qui ont introduit ces méthodes de prédiction. Plusieurs compagnies ont par la suite rapidement mis en oeuvre ces techniques afin de bénéficier des mêmes

gains. Leur usage est avantageux dans plusieurs branches de l'assurance telles que la modélisation du risque de mortalité en assurance-vie, la solvabilité d'une compagnie, la modélisation des réserves, la détection de fraude, etc. Pour de plus ample information concernant l'utilisation de l'apprentissage machine en assurance, on réfère à l'article *Machine Learning in Insurance* de la *Casualty Actuarial Society* (CAS) (Lupton, 2022).

Loi en assurance automobile au Canada

Les lois concernant l'utilisation de caractéristiques protégées varient selon les différentes provinces canadiennes. Certaines provinces sont régulées par des entités privées et d'autres le sont par des entités publiques à travers le gouvernement provincial. La Colombie-Britannique, le Manitoba et la Saskatchewan ont des régulations d'assurance automobile publiques qui se chargent des couvertures de base requises par la loi. Ces trois provinces sont respectivement mandatées par l'*Insurance Corporation of British Columbia* (ICBC), *Manitoba Public Insurance* (MPI) et *Saskatchewan General Insurance* (SGI) qui sont soumis à la Charte canadienne des droits et libertés et doivent suivre les règlements provinciaux en matière d'assurance (West, 2013). La Saskatchewan, par contre, peut offrir des couvertures supplémentaires par le biais d'assureurs privés (Co-operators, 2023). Ces organisations sont les premières à opter pour une tarification qui ne tient pas compte de certains attributs sensibles. En effet, l'ICBC n'utilise pas l'âge, le sexe ou l'état civil dans la tarification des primes en assurance automobile (Insurance Corporation of British Columbia, 2012). En Saskatchewan et au Manitoba, l'utilisation des variables âge et sexe est aussi interdite (SGI, 2023) (Milenkovic, 2021).

Au Nouveau-Brunswick et à Terre-Neuve et Labrador, quoique l'assurance soit

régulée par des organisations privées, l'usage de l'âge, du sexe et de l'état civil pour la tarification automobile est aussi interdit par la loi (Finance and Treasury Board, 2004) (Board of commissioners of public utilities, 2019). La tarification est unisexe et est basée sur d'autres facteurs comme la marque et le modèle de la voiture. Ceci a été instauré dans le but d'avoir une classification plus inclusive et exempte de discrimination.

L'assurance automobile en Alberta et en Ontario est administrée par des régimes d'assurance privés. Les lois dans ces deux provinces n'interdisent pas l'emploi des trois attributs sensibles mentionnés précédemment dans la tarification (Automobile Insurance Rate Board, 2022) (Financial Services Regulatory Authority of Ontario, 2023).

Au Québec, la Société d'assurance automobile du Québec (SAAQ) est un régime d'assurance public qui couvre les blessures corporelles subies lors d'un accident. L'indemnisation ne tient pas compte de la responsabilité du conducteur. Néanmoins, la SAAQ ne remplace pas l'assurance privée qui est obligatoire et qui est fournie par les compagnies d'assurance indépendantes. Selon la Charte des droits et libertés de la personne du Québec, l'usage de certains attributs sensibles est permis si ceux-ci sont démontrés être représentatifs du risque. En effet, la Charte déclare :

> « Dans un contrat d'assurance ou de rente, un régime d'avantages sociaux, de retraite, de rentes ou d'assurance ou un régime universel de rentes ou d'assurance, une distinction, exclusion ou préférence fondée sur l'âge, le sexe ou l'état civil est réputée non discriminatoire lorsque son utilisation est légitime et que le motif qui la fonde constitue un facteur de détermination de risque, basé sur des données actuarielles. [...] » (Gouvernement du Québec, 1996)

Il peut être compliqué d'appliquer cette clause et de démontrer que l'usage de ces variables est bien représentatif du risque. Grâce à la science actuarielle qui évolue constamment et aux techniques de tarification qui ne cessent de se développer, plusieurs travaux de recherche proposent des moyens de quantifier et de contrôler l'inéquité dans les modèles d'apprentissage statistique. Ces travaux seront discutés dans les sections à venir.

CHAPITRE II

DÉFINITIONS

2.1 Définitions des concepts d'équité en actuariat

L'assurance étant une industrie où la discrimination des assurés est une pratique nécessaire pour la solvabilité de la compagnie, il est important de définir les différents concepts d'équité et ceux qui s'appliquent dans l'industrie. Premièrement, on retrouve le *disparate impact* qui, souvent, réfère à la discrimination involontaire (Seiner, 2006) (Paetzold et Willborn, 1996). Le *disparate impact* ou l'*adverse impact* est l'effet qu'une variable sensible a sur les prédictions faites pour les classes protégées, quoique son utilisation ne soit pas faite de manière discriminatoire. Cet impact doit être négatif, comme une prime plus élevée pour les individus appartenant à un certain groupe par exemple, pour être considéré comme un *disparate impact*. De plus, il faut démontrer qu'il existe une méthode alternative pratique qui puisse remplacer la méthode existante qui cause cet impact. En raison de cette dernière condition, aucun assureur n'a été coupable de *disparate impact* à ce jour (Chibanda, 2022). La discrimination involontaire par le biais de variables proxy peut également être présente en assurance, mais dans ce cas on ne réfère plus à un *disparate impact*.

Deuxièmement, on a également le concept de *disparate treatment* qui diffère du *disparate impact*. Le *disparate treatment* est plutôt perçu comme étant de la dis-

crimination intentionnelle (Seiner, 2006) (Paetzold et Willborn, 1996). Ainsi, les individus qui appartiennent à un groupe minoritaire sont délibérément traités défavorablement. Par exemple, la différenciation volontaire sur une classe protégée entre les demandeurs d'emploi constitue un *disparate treatment* à l'égard des personnes appartenant à un groupe minoritaire (Barocas et Selbst, 2016). La discrimination volontaire peut également être faite par le biais de variables proxy et un exemple de cette pratique est le *redlining* qui a été discuté précédemment (Chibanda, 2022).

2.2 Définitions mathématiques

Maintes méthodes ont été proposées dans la littérature afin de mesurer l'injustice dans les algorithmes d'apprentissage machine. Ces dernières diffèrent si on travaille sur un algorithme de classification ou de régression. Plusieurs de ces méthodes seront brièvement présentées.

2.2.1 Algorithme de classification

Dans le cas d'un algorithme de classification où la variable réponse ainsi que l'attribut sensible sont binaires, Grari *et al.* (2022) proposent certains critères d'évaluation pour quantifier l'équité d'un modèle. Le critère de parité démographique, ou de parité statistique, en est un premier exemple. Celui-ci est principalement utilisé dans les scénarios où on cherche à contrôler l'équité de groupe (Dwork *et al.*, 2011). Un algorithme statistique respecte le critère de parité démographique et est considéré juste si

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

La prédiction de la variable réponse, $\hat{Y}$, doit être indépendante de la variable sensible, S, ce qui nous permet d'obtenir des probabilités d'un évènement égales pour tout sous-groupe de l'attribut sensible (Grari *et al.*, 2022). Ce critère peut être appliqué sous forme de diverses mesures afin d'évaluer l'impact disparate du modèle sur les groupes minoritaires. Par exemple, il est possible de l'appliquer en matière d'un ratio tel que (Grari *et al.*, 2022)

$$p\text{-}rule : \min\left\{\frac{\mathbb{P}(\hat{Y}=1|S=0)}{\mathbb{P}(\hat{Y}=1|S=1)}, \frac{\mathbb{P}(\hat{Y}=1|S=1)}{\mathbb{P}(\hat{Y}=1|S=0)}\right\}. \tag{2.1}$$

L'équité du modèle est ensuite déterminée selon un seuil p. Un seuil de 100% constitue un algorithme totalement équitable en ce qui a trait à la parité statistique. Cela indiquerait que les probabilités prédites pour un évènement sont égales, indépendamment du sous-groupe auquel un individu appartient, et cela respecterait le critère initial. Dans les processus d'embauche, par exemple, le seuil minimal p que le ratio doit atteindre selon la loi afin de minimiser le *disparate impact* est de 80% (Feldman *et al.*, 2015). Si ce dernier est inférieur à 80%, il y a discrimination sur une sous-classe de l'attribut sensible, ce qui indiquerait qu'une certaine sous-classe a plus de chance d'être embauchée qu'une autre.

De façon plus générale, on peut réécrire le ratio de parité démographique comme suit :

$$\frac{\max_{s\in S}\mathbb{E}[(\hat{Y}|S=s)]}{\min_{s\in S}\mathbb{E}[(\hat{Y}|S=s)]} \leq \epsilon. \tag{2.2}$$

Pour quantifier l'équité d'un modèle, une seconde mesure existe qui évalue le *disparate impact* et qui est issue de la définition de la parité démographique également. Elle peut se trouver sous forme d'une différence absolue (Feldman *et al.*, 2015) (Grari *et al.*, 2022)

$$DI : |\mathbb{P}(\hat{Y} = 1|S = 1) - \mathbb{P}(\hat{Y} = 1|S = 0)|. \tag{2.3}$$

Plus cette différence est petite, plus le modèle est considéré juste et équitable.

Un deuxième critère de quantification de l'équité est celui des cotes égalisées (*equalized odds*). Les cotes égalisées prennent en compte les taux de faux positifs et faux négatifs. Selon la définition de ce critère, les conditions suivantes doivent être respectées (Grari *et al.*, 2022) :

$$\mathbb{P}(\hat{Y} = 1|Y = 0, S = 0) = \mathbb{P}(\hat{Y} = 1|Y = 0, S = 1) \tag{2.4}$$

$$\mathbb{P}(\hat{Y} = 0|Y = 1, S = 1) = \mathbb{P}(\hat{Y} = 0|Y = 1, S = 0). \tag{2.5}$$

Dans le cas d'une variable sensible binaire, un modèle est dit juste au sens des cotes égalisées si la différence en valeur absolue entre le taux de faux positifs de la première sous-classe de S et le taux de faux positifs de la seconde sous-classe de S, ainsi que la différence en valeur absolue entre le taux de faux négatifs pour la première sous-classe de S et le taux de faux négatifs pour la seconde sous-classe de S sont minimisées (Zafar *et al.*, 2017). Entre autres, les mesures qui sont calculées sont les suivantes :

$$D_{FPR} : \left|\mathbb{P}(\hat{Y} = 1|Y = 0, S = 0) - \mathbb{P}(\hat{Y} = 1|Y = 0, S = 1)\right| \tag{2.6}$$

$$D_{FNR} : \left|\mathbb{P}(\hat{Y} = 0|Y = 1, S = 1) - \mathbb{P}(\hat{Y} = 0|Y = 1, S = 0)\right|. \tag{2.7}$$

Si le taux de faux positifs et le taux de faux négatifs diffèrent largement entre les deux sous-groupes de l'attribut sensible, cela signifie que de l'inéquité existe dans le modèle.

Troisièmement, l'égalité d'opportunité est un critère qui est moins contraignant que celui des cotes égalisées et évalue l'équité d'un modèle selon la disparité entre les taux de vrais positifs dans les sous-groupes de la classe protégée (Hardt *et al.*, 2016). L'équation à respecter selon ce critère est :

$$\mathbb{P}(\hat{Y} = 1|Y = 1, S = 0) = \mathbb{P}(\hat{Y} = 1|Y = 1, S = 1). \tag{2.8}$$

L'équation 2.8 a pour but de donner des chances égales à tous les individus d'une certaine classe protégée et cherche à ce que les taux de vrais positifs soient égaux. L'égalité d'opportunité, comme le nom le dit, évalue l'équité principalement sur le fait que tous les individus aient une chance égale d'avoir une opportunité (une prédiction $\hat{Y} = 1$) quelle qu'elle soit (Hardt *et al.*, 2016).

Dans le cas où la variable sensible contient plus de deux modalités, le ratio de parité démographique présenté à l'équation 2.2 peut être mis en oeuvre. Dans le cas où la variable réponse a aussi plusieurs modalités, une mesure inspirée par l'égalité d'opportunité est proposée par Shen *et al.* (2022). Elle permet de minimiser la somme de la différence absolue entre la perte moyenne calculée sur les instances appartenant à y et à s et la perte moyenne des instances appartenant à y, indépendamment de l'attribut sensible. La somme est effectuée sur toutes les modalités de Y et de S. Cette mesure d'équité est utilisée dans les algorithmes de classification en classes multiples. Pour plus de détails, on réfère à l'article *Optimising Equal Opportunity Fairness in Model Training* (Shen *et al.*, 2022).

2.2.2 Algorithme de régression

Dans les algorithmes de régression, la quantification de l'équité du modèle diffère de celle faite dans les algorithmes de classification. On peut retrouver les mêmes

concepts d'équité tels que la parité démographique et les cotes égalisées, mais avec des mesures différentes qui prennent en compte une variable réponse et une variable sensible continues.

L'idée principale sur laquelle les mesures d'équité dans les algorithmes de classification se basent est de parvenir à faire des prédicitions $\hat{Y}$ indépendantes de l'attribut sensible S. Dans un scénario continu, on procède avec le même principe. Le coefficient de corrélation d'Hirschfeld Gebelein-Rényi (HGR) est une mesure qui évalue l'indépendance entre deux variables continues (Mary *et al.*, 2019). En effet, le coefficient HGR est évalué comme suit (Grari *et al.*, 2022) (Mary *et al.*, 2019) :

$$HGR(U, V) = sup_{f,g}\ \rho(f(U), g(V)), \tag{2.9}$$

où ρ est le coefficient de corrélation de Pearson et les fonctions $f()$ et $g()$ sont deux fonctions mesurables avec $\mathbb{E}[f^2(U)], \mathbb{E}[g^2(V)] < \infty$. L'équation 2.9 évalue le supremum de la corrélation de Pearson entre les fonctions de U et les fonctions de V, où U et V sont deux variables aléatoires continues. Le coefficient se situe entre [0, 1]. Une valeur de 0 indique que U et V sont indépendantes, alors qu'une valeur de 1 indique qu'elles ne le sont pas (Mary *et al.*, 2019). La mise en oeuvre de cette mesure dans une fonction objective pénalisée est abordée plus en détails dans l'article « *Fairness-Aware Learning for Continuous Attributes and Treatments* » (Mary *et al.*, 2019).

Un second test qui permet d'évaluer l'équité d'un modèle est celui de la parité d'erreur (*error parity*). Ce dernier évalue si les erreurs des prédictions faites par un algorithme sont distribuées de manière analogue entre les sous-groupes d'un attribut sensible, et, si c'est bien le cas, le modèle est considéré équitable selon cette mesure (Gursoy et Kakadiaris, 2022). En d'autres mots, ceci implique que

l'erreur, r, doit être indépendante de l'attribut sensible, s :

$$r \perp\!\!\!\perp s. \tag{2.10}$$

L'erreur, r, dans l'équation 2.10 est une fonction de $(y, \hat{y})$ et le choix de cette fonction peut varier selon l'algorithme avec lequel on travaille. Quelques exemples sont la différence entre la vraie valeur et la prédiction, $y - \hat{y}$, la différence absolue entre ces dernières, $|y - \hat{y}|$, ou la différence aux carrées,$(y - \hat{y})^2$ (Gursoy et Kakadiaris, 2022).

Par ailleurs, la divergence entre deux distributions peut être calculée avec des tests de qualité d'ajustement (*goodness-of-fit tests*) comme le test de Kolmogorov-Smirnov ou celui du khi-carré. Le choix du test dépend surtout de la nature des distributions qu'on cherche à comparer. Le test de Kolmogorov-Smirnov, par exemple, est moins sensible à la queue de la distribution et est moins apte à capter la qualité d'ajustement si les distributions ont une lourde queue (Gursoy et Kakadiaris, 2022). Pour qu'un algorithme de décision automatisée soit considéré équitable au sens de la parité d'erreur, il faut que le test de qualité d'ajustement conclut que les distributions des erreurs de tous les sous-groupes de l'attribut sensible sont similaires pour un seuil de confiance donné.

2.3 Divergence Kullback-Leibler

La divergence de Kullback-Leibler, aussi appelée entropie relative, est une mesure de qualité d'ajustement qui compare deux distributions, en prenant la vraie distribution des données, dénotée par $P(x)$, qu'on approxime en ajustant une distribution théorique $Q(x)$ sur les mêmes données. La divergence de Kullback-Leibler nous permet de quantifier l'information perdue avec un tel modèle. Elle

mesure la divergence entre une distribution de probabilité $P(x)$ et une distribution ajustée $Q(x)$. La divergence de Kullback-Leibler, dans un contexte discret, est définie avec les deux équations suivantes qui représentent la direction *avant* et *arrière* respectivement :

$$D_{KL}(P||Q) = \sum P(x) log\left(\frac{P(x)}{Q(x)}\right) \tag{2.11}$$

$$D_{KL}(Q||P) = \sum Q(x) log\left(\frac{Q(x)}{P(x)}\right). \tag{2.12}$$

Si les distributions P et Q sont continues, l'entropie relative est évaluée comme suit :

$$D_{KL}(P||Q) = \int p(x) log\left(\frac{p(x)}{q(x)}\right) dx \tag{2.13}$$

$$D_{KL}(Q||P) = \int q(x) log\left(\frac{q(x)}{p(x)}\right) dx \tag{2.14}$$

avec $p(x)$ et $q(x)$ étant respectivement la fonction de densité de probabilité réelle des données et la fonction de densité de probabilité ajustée à ceux-ci. La divergence de Kullback-Leibler n'est pas une mesure de distance, mais plutôt de divergence entre deux distributions, ce qui fait en sorte que les équations 2.11 et 2.12 ne sont pas symétriques, et donc que $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. Il en va de même pour les équations 2.13 et 2.14 (Bishop, 2006). Elles peuvent prendre des valeurs entre [0, ∞]. Une divergence de 0 indique qu'il n'y a pas de perte d'information en ajustant la distribution Q pour approximer P, et que $Q(x) = P(x)$ pour tout $x \in \mathcal{X}$.

L'entropie relative a beaucoup été utilisée en tant que fonction d'optimisation dans les algorithmes d'apprentissage machine, plus précisément dans les modèles d'apprentissage profond qui entrainent des réseaux neurones artificiels par des autos-encodeurs variationnels (Kingma et Welling, 2022) ou les modèles de *Deep*

Neural Networks (DNNs) (Tishby et Zaslavsky, 2015). À noter que la divergence de Kullback-Leibler peut être mise en oeuvre en tant que fonction objectif dans les deux sens tels que présentés dans les équations 2.11 et 2.12 ainsi que dans les équations 2.13 et 2.14. Puisque cette dernière n'est pas symétrique, la fonction $Q(x)$ résultante ne sera pas la même dans les deux cas. Le choix de la direction dans laquelle elle est évaluée est définie par la connaissance de la vraie distribution $P(x)$. Si elle est connue, la direction arrière peut être appliquée. Si on a un échantillon de données de la vraie distribution, la direction avant est favorisée (Ghosh, 2018).

Dans le présent mémoire, la divergence de Kullback-Leibler sera mise en oeuvre en tant que contrainte d'équité, additionnellement à la fonction d'optimisation du modèle, dans le but de contrôler le *disparate impact* causé par les attributs sensibles dans les données d'entrainement. Ceci sera discuté plus en détails ultérieurement.

2.4 Intervention possible

Afin de rendre un modèle statistique équitable, il y a trois moments où il est possible d'intervenir dans la construction de l'algorithme : avant l'entrainement, la méthode qu'on appelle le *preprocessing*, pendant l'entrainement, le *in-processing* ou après l'entrainement, le *post-processing*.

2.4.1 *Preprocessing*

Le *preprocessing* consiste à modifier la base de données avec laquelle on travaille afin d'éliminer les biais qui peuvent s'y trouver initialement. Il existe plusieurs définitions et mesures de biais. L'Institut canadien des actuaires définit le biais en assurance comme suit :

> « S'entend de biais dans la tarification des assurances IARD toute situation où les résultats des modèles de tarification sont systématiquement moins favorables pour les personnes d'un groupe particulier et où il n'y a pas de différence pertinente entre les groupes justifiant un tel écart de primes ou de taux » (Institut canadien des actuaires, 2023).

Dans la méthode du *preprocessing*, la gestion du biais débute avant même la construction du modèle prédictif. Le contrôle du biais est géré directement dans la base de données. Un exemple de biais qu'on peut vouloir éliminer des données est le biais d'échantillonnage qui signifie que certaines populations risquent d'être mal représentées à cause d'un manque de données à leur égard durant la collecte d'informations, ce qui rend le modèle moins performant pour ces groupes minoritaires et cela peut mener à une discrimination envers eux dans les prédictions du modèle (Mehrabi *et al.*, 2022) (Chouldechova et Roth, 2018).

Un autre exemple est le biais historique. Celui-ci peut s'infiltrer dans les données qui sont utilisées dans le modèle statistique et, par conséquent, renforcer un stéréotype discriminatoire qui sera reflété dans les prédictions (Suresh et Guttag, 2021) (Mehrabi *et al.*, 2022). Plusieurs autres types de biais peuvent être perçus dans les données et Ninareh Mehrabi et al. en font une description complète et détaillée (Mehrabi *et al.*, 2022). La méthode du *preprocessing* sert à rectifier ces types de biais et plusieurs autres afin d'être en mesure d'entrainer l'algorithme avec des données impartiales.

Une manière de faire est de normaliser la variable réponse par rapport à la classe protégée correspondante afin que les données soit bien réparties parmi les sous-groupes. Le premier algorithme présenté par Mohamed et Schuller, qui est intitulé *FaiReg*, solutionne le biais d'étiquetage, un biais intrinsèque dans la base de données utilisées (Mohamed et Schuller, 2022). La seconde méthode de *preprocessing* présentée est, *FaiRegH*, qui est une méthode hybride entre *FaiReg* et l'équili-

brage des données. L'équilibrage des données consiste à attribuer différents poids à différentes observations de l'attribut sensible. De plus, cette méthode vise à suréchantilloner de manière aléatoire les catégories qui contiennent peu d'observations dans le but d'avoir des sous-classes de l'attribut sensible de même taille et qui suivent des distributions similaires (Yan *et al.*, 2020). De même, la technique d'équilibrage des données peut servir à sous-échantillonner certains groupes afin de rendre leur distribution similaire à celles des autres sous-groupes. Cependant, le sous-échantillonage n'est pas favorisé puisqu'il risque d'éliminer des observations et de l'information fondamentales à l'entrainement du modèle (Mohamed et Schuller, 2022). L'approche *FaiRegH* vise à rectifier simultanément les biais d'étiquetage et d'échantillonnage.

2.4.2 *In-processing*

Le deuxième moment d'intervention possible est la méthode du *in-processing* qui consiste à mitiger la discrimination durant le processus d'entrainement, entre autre en ajoutant une contrainte d'équité ou en modifiant la fonction d'optimisation. Les travaux de Berk *et al.* (2017) mettent en oeuvre cette technique. Leur approche consiste à rajouter une contrainte d'équité à la fonction de perte pour la précision. La fonction d'optimisation initialement proposée est la suivante :

$$\ell_{\mathcal{P}}(w, S) + \lambda f_{\mathcal{P}}(w, S), \tag{2.15}$$

où $\ell_{\mathcal{P}}(w, S)$ est la fonction de perte pour la précision en fonction du modèle w et un échantillon de données d'entrainement noté S et $f_{\mathcal{P}}(w, S)$ est la contrainte d'équité ajoutée. Un poids λ est également assigné à la contrainte d'équité afin de contrôler le niveau d'équité qu'on souhaite avoir dans le modèle. Ce poids est essentiel puisqu'il faut considérer qu'il y a un compromis entre la précision et le

niveau d'équité du modèle (Bian et Zhang, 2023). Un avantage avec la méthode d'intervention *in-processing* est le contrôle de ce compromis précision-équité.

La fonction $\ell_{\mathcal{P}}(w, S)$ est convenu d'être l'erreur quadratique moyenne (MSE) pour les problèmes de régression linéaire ou le *Log-Loss* dans le cadre d'une régression logistique. Les fonctions $f_{\mathcal{P}}(w, S)$ choisies dans l'article de Berk *et al.* (2017) tiennent compte de la définition d'équité que les auteurs ont décidé d'adopter. Un modèle est considéré équitable si un modèle traite de façon similaire deux observations similaires, chacune appartenant à un sous-échantillon différent, soit S_1 ou S_2, deux sous-groupes de S. Pour l'équité individuelle, la fonction suivante a été proposée :

$$f_1(w, S) = \frac{1}{|S_1||S_2|} \sum_{\substack{(x_i,y_i)\in S_1 \\ (x_j,y_j)\in S_2}} d(y_i, y_j)(wx_i - wx_j)^2. \tag{2.16}$$

La fonction d est choisie de manière à normaliser les données réelles afin d'obtenir une moyenne de 0 et une variance de 1. Pour les problèmes de régression, la fonction $d(y_i, y_j) = e^{-(y_i - y_j)^2}$ et, dans les problèmes de classification binaire, on retrouve $d(y_i, y_j) = \mathbb{1}[y_i = y_j]$. Cette fonction de normalisation n'influence pas le choix du modèle w puisqu'elle est considérée comme une constante par rapport à ce dernier (Berk *et al.*, 2017). La contrainte d'équité ajoutée dans un cadre d'équité de groupe est la suivante :

$$f_2(w, S) = \left(\frac{1}{|S_1||S_2|} \sum_{\substack{(x_i,y_i)\in S_1 \\ (x_j,y_j)\in S_2}} d(y_i, y_j)(wx_i - wx_j) \right)^2. \tag{2.17}$$

De plus, la régularisation L2 a été ajoutée dans le problème d'optimisation et a été pondérée par γ, une fonction de λ, qui est évaluée par validation croisée. La perte résultante à optimiser devient donc :

$$argmin_w(\ell_{\mathcal{P}}(w,S) + \lambda f_{\mathcal{P}}(w,S) + \gamma||w||_2). \tag{2.18}$$

Une méthode semblable sera mise en oeuvre dans le cadre de ce mémoire en incorporant la divergence de Kullback-Leibler, discutée précédemment, en tant que contrainte d'équité. Il existe maintes méthodes de *in-processing*. Pérez-Suay *et al.* (2017) en présente une qui diffère de celle de Berk *et al.* (2017) (Pérez-Suay *et al.*, 2017).

2.4.3 *Post-processing*

Le dernier moment d'intervention est après l'entrainement, une fois que les prédictions sont faites par l'algorithme. Cette méthode d'intervention se nomme le *post-processing* et celle-ci cherche à rendre les prédictions le plus équitable possible en les altérant selon différentes méthodes. Un algorithme de modification des prédictions par une combinaison linéaire qui implémente l'équité selon la définition de la parité démographique est proposé comme une solution parmi tant d'autres pour mitiger la discrimination (Chzhen *et al.*, 2020). Pour une classe protégée binaire, on peut réécrire la combinaison linéaire comme suit :

$$g^*(x,s) = p_s f^*(x,s) + (1-p_s)t^*(x,s). \tag{2.19}$$

L'équation précédente montre que g^* est la nouvelle prédiction équitable qui est une combinaison de f^* et t^*, pondéré par p_s, la fréquence du groupe s. La distribution $f^*(x,s)$ est la distribution obtenue par entrainement du modèle. Dans un cas simplifié où la classe protégée est binaire, les distributions de $f^*(x,1)$ et $f^*(x,2)$ ne sont pas nécessairement similaires ce qui peut créer de la disparité dans les prédictions pour ces 2 sous-classes. Un ajustement de la prédiction est donc

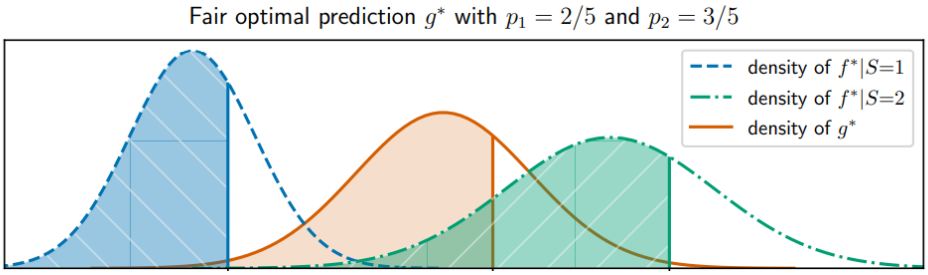


FIGURE 2.1 – Densité des sous-classes d'un attribut sensible S binaire (Chzhen *et al.*, 2020)

effectué en prenant en compte la valeur de $t^*(x, s)$ obtenue en égalisant l'aire de la courbe bleue de la Figure 2.1 avec l'aire de la courbe verte (Chzhen *et al.*, 2020).

Entre autres $t^*(x, s)$ peut être obtenue en résolvant l'équation suivante pour un cas binaire :

$$\mathbb{P}(f^*(X, S) \leq f^*(x, 1)|S = 1) = \mathbb{P}(f^*(X, S) \leq t^*(x, 1)|S = 2). \tag{2.20}$$

L'idée est de choisir la valeur $t^*(x, 1)$ dans la densité de $f^*|S = 2$ qui a une probabilité cumulative égale à la probabilité cumulative de l'évènement $f^*(x, 1)$ dans la densité de $f^*|S = 1$. Ensuite, une prédiction équitable, $g^*(x, s)$ est calculée grâce à l'équation 2.19. Finalement, l'équité de g^* est quantifiée avec la mesure du *disparate impact* (équation 2.3) d'après la définition de la parité démographique. On évalue donc (Chzhen *et al.*, 2020) :

$$sup_{t\in\mathbb{R}}\Big|\mathbb{P}(g(X, S) \leq t|S = s) - \mathbb{P}(g(X, S) \leq t|S = s')\Big| = 0. \tag{2.21}$$

Dans les cas où l'entrainement d'un algorithme est coûteux, intervenir après l'entrainement en utilisant la méthode *post-processing* peut être avantageux. Cependant, le moment d'intervention, tout comme la définition d'équité utilisée, dépend

du modèle et des données disponibles. L'équité est un problème spécifique au domaine et il est crucial de bien comprendre la modélisation qui est mise en place et les données avec lesquelles on travaille afin de pouvoir appliquer les ajustements nécessaires pour obtenir des prédictions justes.

CHAPITRE III

MODÈLES

Les différents modèles utilisés dans le cadre de cette recherche sont présentés dans le présent chapitre. On retrouve le modèle linéaire généralisé (MLG) Tweedie pour la modélisation de la perte encourue, le MLG Bernoulli qui modélise la probabilité de survenance d'un sinistre pour chaque assuré et le MLG Gamma qui modélise la sévérité d'un sinistre. On retrouve également les différents modèles de Boosting : le Boosting Tweedie, avec et sans contrainte d'équité, et le Boosting de classification binaire, avec et sans contrainte d'équité.

3.1 Théorie des modèles linéaires généralisés (MLG)

Les modèles linéaires généralisés sont des modèles statistiques prédictifs qui permettent de relier, par l'entremise d'une fonction lien, une variable réponse Y à un prédicteur linéaire, $\boldsymbol{X\beta}$, $\boldsymbol{X}$, de dimensions $n \times (p+1)$, étant la matrice des variables explicatives du modèle, avec n le nombre d'observations et p le nombre de variables explicatives, et $\boldsymbol{\beta}$, de dimensions $(p+1) \times 1$, le vecteur des coefficients. Les prédictions faites par le modèle peuvent s'écrire comme suit

$$\mathbb{E}(Y|\boldsymbol{X}) = \hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{3.1}$$

L'espérance conditionnelle de Y, notée μ, est aussi égale à l'inverse de la fonction lien canonique g évalué en $\boldsymbol{X\beta}$. Le choix de la fonction lien va souvent avec le choix de la famille de distribution choisie pour la modélisation de Y. De plus, dans les modèles linéaires généralisés, la variance peut être exprimée comme une fonction de variance V telle que

$$Var(Y|\boldsymbol{X}) = V(\mu) = V(g^{-1}(\boldsymbol{X\beta})). \tag{3.2}$$

La variance n'est donc pas une constante comme dans les hypothèses des modèles de régression linéaire classiques, mais est plutôt une fonction des variables explicatives. Une seconde caractéristique propre aux MLG est que la distribution de Y doit appartenir à la famille exponentielle linéaire. Dans cette famille, les fonctions de densité de probabilité peuvent être écrites sous cette forme

$$f_Y(y) = c(y, \phi)exp\Big(\frac{y\theta - a(\theta)}{\phi}\Big), \tag{3.3}$$

où θ est le paramètre canonique ($g(\mu) = \theta$), ϕ est le paramètre de dispersion, $c()$ est la mesure de base et $a()$ est la log-partition (Pigeon, 2022) (Clark et Thayer, 2004).

L'estimation des paramètres d'un modèle linéaire généralisé est généralement faite par maximum de vraisemblance même si d'autres critères d'estimation existent comme le critère des moindres carrées. On cherche à maximiser la vraisemblance ou à minimiser la log-vraisemblance négative de la famille exponentielle choisie pour la modélisation. La log-vraisemblance est souvent utilisée pour faciliter les manipulations et les calculs.

Dans le cadre de cette recherche, deux modèles linéaires généralisés de tarification automobile sont mis en oeuvre. Le premier suppose que Y, la perte totale, suit

une distribution de la famille Tweedie et sera présenté à la sous-section 3.1.1. Le deuxième modélise la survenance d'un sinistre et la sévérité de ce dernier séparément. Ceci résulte en la construction de deux MLG distincts. Le premier, où la variable réponse suit une distribution Binomiale, prédit la probabilité d'occurrence d'un sinistre pour l'assuré i. Le deuxième, où la variable réponse suit une distribution Gamma, modélise la sévérité sachant qu'un sinistre a eu lieu. Cette seconde approche sera détaillée à la sous-section 3.1.2.

3.1.1 Modèle linéaire généralisé Tweedie

Afin de modéliser la perte totale des assurés, la distribution Tweedie est souvent utilisée puisque celle-ci a la capacité de prendre en compte le point de masse à 0 présent dans les données en assurance. En effet, ce phénomène est souvent perçu dans l'industrie de l'assurance puisque la plupart des assurés n'ont pas de réclamation. La distribution Tweedie est aussi souvent référée comme étant une distribution Poisson-Gamma composée, avec un paramètre ρ (*power parameter*) qui est déterminé par analyse du profil de vraisemblance. Plusieurs modèles avec des valeurs différentes pour le paramètre ρ sont ajustés et la valeur de ρ qui maximimise la vraisemblance est retenue pour l'ajustement du modèle final. Il est possible de retrouver certaines distributions connues en fixant ρ à une valeur précise. Par exemple, avec $\rho = 1$, on retrouve une distribution Poisson, alors qu'avec $\rho = 2$, on retrouve une distribution Gamma. Pour une distribution composée Poisson-Gamma, le paramètre ρ doit se situer entre 1 et 2. De plus, pour les distributions de la famille Tweedie, l'espérance et la variance sont reliées grâce au paramètre ρ et au paramètre de dispersion ϕ

$$E(Y) = \mu \tag{3.4}$$

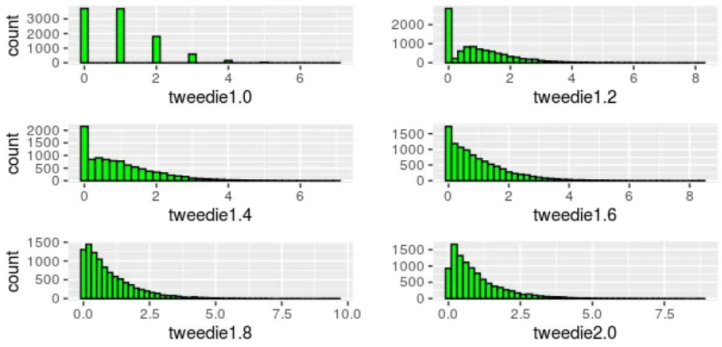


FIGURE 3.1 – Répartition des fréquences appartenant à la famille Tweedie selon différente valeur de ρ (Tiwari, 2020).

$$Var(Y) = \phi\mu^{\rho}. \tag{3.5}$$

Un graphique de la répartition des fréquences appartenant à la distribution Tweedie avec différentes valeurs de ρ est représentée à la Figure 3.1 alors que tous les autres paramètres restent inchangés. Il est important de noter que la fonction de densité de probabilité (f.d.p.) Tweedie n'a pas de forme fermée explicite, mais on sait qu'elle appartient à la famille exponentielle linéaire. Par conséquent, on peut l'approximer en l'écrivant sous la forme présentée à l'équation 3.3

$$f_Y(y) = c(y, \phi, \rho)exp\Big\{\frac{1}{\phi}\Big(y\frac{\mu^{1-\rho}}{1-\rho} - \frac{\mu^{2-\rho}}{2-\rho}\Big)\Big\}. \tag{3.6}$$

Cette forme est applicable lorsque $1 < \rho < 2$, le cas dont il est question dans ce modèle. La fonction $c(y, \phi, \rho)$ n'est pas connue, mais elle considérée comme étant une constante dans le problème d'optimisation et peut donc être ignorée. Il en est de même pour la constante $\frac{1}{\phi}$. En utilisant cette approximation de la f.d.p., il est possible d'écrire l'équation d'optimisation qui est la log-vraisemblance négative suivante :

$$L = -\sum_i y_i \frac{\hat{y_i}^{1-\rho}}{1-\rho} + \frac{\hat{y_i}^{2-\rho}}{2-\rho}, \tag{3.7}$$

où les prédictions de l'assuré i, $\hat{y_i}$, remplacent le paramètre μ dans l'équation 3.6 (Yang *et al.*, 2016a). Cette fonction est donc minimisée pour l'ajustement du premier MLG Tweedie.

L'équité du MLG Tweedie est comparée à celle du modèle de Boosting Tweedie sans contrainte d'équité, ainsi qu'à celle du Boosting Tweedie en ajoutant la divergence de Kullback-Leibler comme contrainte d'équité. Ces modèles seront discutés dans les sections à venir.

3.1.2 Modèle linéaire généralisé Bernoulli-Gamma

Pour le modèle prédictif d'occurrence-sévérité, le premier MLG qui a été fait en est un qui suppose une distribution Binomiale, plus précisément une distribution Bernoulli, puisque la variable réponse, Z, est binaire (0-1) prédisant la probabilité de survenance d'un sinistre, soit $P(Z = 1|\boldsymbol{X})$. Ce modèle est comparé à un modèle de Boosting de classification binaire, prédisant également la probabilité de sinistre pour l'assuré i, ainsi qu'à un modèle de Boosting incorporant la divergence de Kullback-Leibler comme métrique d'équité.

La variable réponse modélisée, Z, est égale à 1 s'il y a occurence d'un sinistre et elle est égale à 0 sinon. Un modèle de famille Binomiale avec un lien logistique est donc mis en oeuvre. Le lien entre l'espérance μ et l'équation linéaire entre les prédicteurs $\boldsymbol{X}$ et les coefficients $\boldsymbol{\beta}$ équivaut alors à

$$g(\mu) = log\Big(\frac{\mu}{1-\mu}\Big) = \boldsymbol{X\beta}. \tag{3.8}$$

Ceci nous donne des prédictions sous la transformation suivante

$$\hat{\mu} = P(Z = 0|\boldsymbol{X}) = \frac{e^{-\boldsymbol{X}\hat{\boldsymbol{\beta}}}}{1 + e^{-\boldsymbol{X}\hat{\boldsymbol{\beta}}}}, \tag{3.9}$$

$$\hat{\mu} = P(Z = 1|\boldsymbol{X}) = \frac{1}{1 + e^{-\boldsymbol{X}\hat{\boldsymbol{\beta}}}}. \tag{3.10}$$

Le deuxième MLG est celui qui modélise la sévérité sachant qu'un sinistre a eu lieu. Seuls les assurés qui ont subi un sinistre sont pris en compte pour l'entrainement de ce modèle. On a posé comme hypothèse que la sévérité suit une distribution Gamma et la fonction lien appliquée dans ce cas-ci est le lien logarithmique :

$$g(\mu) = log(\mu) = \boldsymbol{X}\boldsymbol{\beta}. \tag{3.11}$$

Ce qui nous donne des prédictions telles que

$$\hat{\mu} = e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}}. \tag{3.12}$$

Pour obtenir la prime pure (PP) d'un assuré donné, la prédiction de probabilité de survenance (Z) pour l'assuré i est multipliée avec la prédiction de sévérité (L) du même assuré

$$\mathbb{E}[PP] = \mathbb{E}[Z] \times \mathbb{E}[L]. \tag{3.13}$$

3.2 Théorie des algorithmes de Boosting

Les modèles de Boosting sont des algorithmes prédictifs qui tentent de minimiser l'erreur de prédiction grâce à un entrainement séquentiel. La prédiction finale de

ces algorithmes est faite par l'entremise de plusieurs apprenants faibles (*weak learner*) qui forment un apprenant fort (*strong learner*). Le modèle abordé dans cette section emploie une séquence d'arbres de décision comme apprenant faible. Un sous-échantillon aléatoire de données est initialement choisi pour mettre en oeuvre un premier arbre de décision. La proportion de sous-échantillonage est contrôlée par un hyperparamètre du modèle de Boosting. Un sous-échantillon de variables explicatives est également choisi pour l'entrainement selon un hyperparamètre qui définit le nombre de variables à utiliser pour chaque apprenant faible. Les hyperparamètres et leur ajustement sont discutés plus en détails au chapitre suivant. Par la suite, d'autres arbres sont ajustés en améliorant les faiblesses de l'arbre précédent. Cette technique est appelée *ensemble learning* et la prédiction finale est faite par la moyenne empirique de tous les apprenants faibles (*weak learners*) (IBM, 2023). Cette méthode d'entrainement améliore le compromis biais-variance présent dans les arbres de décision. En effet, les arbres de décision peuvent facilement être surajustés si l'élagage (*pruning*) de l'arbre n'est pas mis en oeuvre adéquatement, menant ainsi à une haute variance et à un faible biais. Dans le cas de sous-ajustement (lorsque l'arbre de décision est petit), la variance diminue, mais le biais augmente.

Un squelette de l'algorithme de Boosting par descente de gradient est (Pigeon, 2021) (Yang *et al.*, 2016b) (Friedman, 2001)

1. Initialiser la fonction (r) qu'on cherche à minimiser

$$r_0(\boldsymbol{X}) = argmin_{\hat{y}_i \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \hat{y}_i).$$

2. Pour t allant de 1 à T :

   (a) Calculer $U_i = -\frac{\partial L(Y_i, r_{t-1}(\boldsymbol{X}))}{\partial r_{t-1}(\boldsymbol{X})}$, $i = 1, ..., n$.

(b) Ajuster un apprenant faible, $h_t(\boldsymbol{X})$, qui prédit les résidus $U_i, i = 1, ...n$ en fonction des observations $\boldsymbol{X}$.

(c) Ajuster $r_t(\boldsymbol{X}) = r_{t-1}(\boldsymbol{X}) + \alpha h_t(\boldsymbol{X})$.

3. La prédiction finale est $r_T(\boldsymbol{X})$.

On a $L()$ une fonction de perte évaluée sur un échantillon de taille n avec Y_i la variable réponse et $\hat{y}_i$ les prédictions. On a aussi $\boldsymbol{X}$ les variables explicatives et α le taux d'apprentissage. Plusieurs hyperparamètres peuvent également être ajustés au modèle. Ceux-ci optimisent l'entrainement en minimisant la fonction de perte de l'algorithme, tout en empêchant le sur-ajustement du modèle et en décorrélant les arbres entre eux.

3.2.1 Boosting Tweedie sans contrainte d'équité

Le premier modèle de Boosting qui a été mis en oeuvre est celui d'un Boosting Tweedie afin de modéliser la perte encourue des assurés. La fonction d'optimisation pour ce modèle est la même que celle présentée à l'équation 3.7 pour le MLG Tweedie. La mise en oeuvre manuelle de cette fonction d'optimisation est nécessaire pour pouvoir éventuellement incorporer la divergence de Kullback-Leibler dans le modèle de Boosting avec contrainte d'équité. Pour se faire, l'équation 3.7 doit légèrement être modifiée afin de prendre en compte la fonction lien utilisée dans l'algorithme. Dans ce cas-ci la fonction lien est le lien logarithmique, ce qui signifie que les prédictions sont données par l'équation 3.12 et que la même transformation dans l'équation d'optimisation est nécessaire. Cela fait en sorte que la fonction de perte Tweedie devient

$$\begin{aligned} L &= -\sum_i y_i \frac{e^{\hat{\boldsymbol{z}}*(1-\rho)}}{1-\rho} + \frac{e^{\hat{\boldsymbol{z}}*(2-\rho)}}{2-\rho} \\ &= -\sum_i y_i \frac{e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}*(1-\rho)}}{1-\rho} + \frac{e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}*(2-\rho)}}{2-\rho}. \end{aligned} \tag{3.14}$$

Afin de minimiser les paramètres $\boldsymbol{\beta}$ du modèle, le gradient et la hessienne sont calculés

$$\frac{\partial L}{\partial \beta_j} = -\sum_i y_i e^{\hat{\boldsymbol{z}}*(1-\rho)} + e^{\hat{\boldsymbol{z}}*(2-\rho)} \Big(\frac{\partial \hat{\boldsymbol{z}}}{\partial \beta_j} \Big), \tag{3.15}$$

$$\frac{\partial^2 L}{\partial \beta_j \beta_k} = -\sum_i y_i (1-\rho) e^{\hat{\boldsymbol{z}}*(1-\rho)} + (2-\rho) e^{\hat{\boldsymbol{z}}*(2-\rho)} \Big(\frac{\partial^2 \hat{\boldsymbol{z}}}{\partial \beta_j \beta_k} \Big). \tag{3.16}$$

3.2.2 Boosting Bernoulli sans contrainte d'équité

Un modèle de Boosting pour prédire la probabilité d'occurrence d'un sinistre pour chaque assuré a également été mis en oeuvre. Puisque le modèle en est un de classification binaire, la fonction d'optimisation utilisée est la perte logistique telle que

$$\ell = \frac{-1}{n} \sum_{i=1}^{n} (y_i \log(\hat{q}_i) + (1-y_i) \log(1-\hat{q}_i)), \tag{3.17}$$

où $\hat{q}_i$ est défini par

$$\hat{q}_i = P(Z=1|\boldsymbol{X}) = \frac{1}{1+e^{-\hat{\boldsymbol{z}}}} = \frac{1}{1+e^{-\boldsymbol{X}\hat{\boldsymbol{\beta}}}}. \tag{3.18}$$

Ainsi, un modèle dont les paramètres sont obtenus en minimisant la fonction de perte 3.17 est un modèle dans lequel on choisit les paramètres de façon à minimiser la divergence entre la distribution obtenue et la vraie distribution Bernoulli. Le gradient est alors

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} (\hat{q}_i - y_i) \left(\frac{\partial \hat{q}_i}{\partial \beta_j} \right) \tag{3.19}$$

et la hessienne est

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} \hat{q}_i (1 - \hat{q}_i) \left(\frac{\partial^2 \hat{q}_i}{\partial \beta_j \partial \beta_k} \right). \tag{3.20}$$

L'algorithme de Boosting est programmé de manière à optimiser les $\boldsymbol{\beta}$ avec les dérivées partielles plus haut.

Pour le calcul de la prime pure, on multiplie la probabilité de survenance obtenue par l'algorithme de Boosting sans contrainte d'équité avec les prédictions de sévérité faites par le MLG de famille Gamma qui modélise la sévérité. Étant donné que la sévérité d'un sinistre ne dépend pas des caractéristiques d'un assuré, la discrimination n'est pas évaluée sur ce modèle. Les variables sensibles risquent donc d'affecter davantage la prédiction d'occurrence d'un sinistre et de pénaliser les individus appartenant à un groupe sensible en prédisant une probabilité de survenance plus élevée pour ce groupe minoritaire.

3.2.3 Boosting Tweedie avec contrainte d'équité

Comme présenté à la section 2.3, la divergence de Kullback-Leibler sera ajoutée à la fonction objectif en tant que contrainte d'équité. En effet, cette métrique sera ajoutée à l'équation 3.14 pour obtenir la nouvelle fonction objectif à optimiser. On cherche à minimiser cette divergence afin d'avoir des prédictions qui

se rapprochent le plus possible des vraies données, ce qui va permettre d'éviter de faire des prédictions discriminatoires. La divergence KL est égale à 0 lorsque $P(x) = Q(x)$.

Un modèle ajusté sur une base de données d'entrainement quelconque est parfait si, lorsque évalué sur cette même base de données, il suit une distribution uniforme. Par exemple, en prenant $\boldsymbol{X}$, la matrice des variables explicatives dans la base de données d'entrainement et $F^*(.)$, le modèle ajusté sur ces données, on obtient

$$F^*(\boldsymbol{x}) \sim U(0,1). \tag{3.21}$$

Pour l'algorithme de Boosting Tweedie, $F^*(.)$ est une distribution Tweedie avec paramètre ρ et ϕ égaux à ceux approximés par le MLG Tweedie, et μ, les prédictions du modèle de Boosting. L'hypothèse initiale est que les quantiles du modèle ajusté, dans le cas d'un modèle parfait, devraient suivre une densité uniforme et donc minimiser la divergence de Kullback-Leibler. En évaluant la divergence de manière à comparer ces deux distributions telles que

$$D_{KL}(F^*(\boldsymbol{x}), g(v)) \approx 0, \tag{3.22}$$

avec $g(v)$ la fonction de densité de probabilité d'une uniforme, il est possible de quantifier l'équité du modèle de Boosting sans et avec métrique d'équité grâce à la métrique de parité démographique présentée à l'équation 2.2.

En assignant le sexe comme variable sensible, le calcul de l'équation 3.22 doit être effectué sur les prédictions des femmes et sur celles des hommes séparément, ce qui donne

$$D_{KL}^{F}(F^{*}(\boldsymbol{x}|S=F), g(v)), \tag{3.23}$$

$$D_{KL}^{H}(F^{*}(\boldsymbol{x}|S=H), g(v)). \tag{3.24}$$

La comparaison de ces deux valeurs est ensuite faite en appliquant la parité démographique sur la divergence KL

$$\frac{\max_{s\in S} D_{KL}(F^{*}(\boldsymbol{x}|S=s), g(v))}{\min_{s\in S} D_{KL}(F^{*}(\boldsymbol{x}|S=s), g(v))} \leq \epsilon. \tag{3.25}$$

Une valeur de 1 indiquerait une équité parfaite entre les groupes, cependant, il est possible d'établir un seuil ϵ pour lequel l'équité du modèle serait acceptable selon le contexte.

La fonction de perte pour le Boosting Tweedie avec contrainte d'équité est donc

$$L_{D_{KL}} = -\sum_{i} y_i \frac{e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}*(1-\rho)}}{1-\rho} + \frac{e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}*(2-\rho)}}{2-\rho} + \lambda * \Big(D_{KL}^{F}(F^{*}(\boldsymbol{x}), 1) + D_{KL}^{H}(F^{*}(\boldsymbol{x}), 1) \Big). \tag{3.26}$$

Le gradient et la hessienne de la perte Tweedie sont équivalents à l'équation 3.15 et 3.16 et le gradient et la hessienne de la divergence de Kullback-Leibler sont calculés numériquement. L'hyperparamètre λ permet d'assigner un poids à l'équité qu'on désire avoir dans l'algorithme et de contrôler le compromis précision-équité du modèle.

3.2.4 Boosting Bernoulli avec contrainte d'équité

Pour le modèle d'occurrence-sévérité, la divergence de Kullback-Leibler a également été ajoutée au *Log-loss* dans le premier modèle de classification binaire. L'optimisation de la perte logistique peut être interprétée comme l'optimisation de l'entropie croisée. En effet, l'entropie croisée est définie comme suit

$$H(P,Q) = -\sum P(x) * log(Q(x)), \tag{3.27}$$

avec $P(x)$ la vraie distribution des données et $Q(x)$ le modèle ajusté. Il est aussi possible de montrer que l'optimisation de la divergence de Kullback-Leibler, dans un contexte discret et binaire, est équivalente à l'optimisation de la perte logistique. Pour ce faire, il suffit de réécrire la divergence en terme de l'entropie croisée et de l'entropie. En définissant l'entropie ainsi

$$S_p = -\sum P(x) * log(P(x)), \tag{3.28}$$

la divergence de Kullback-Leibler peut donc s'écrire comme suit

$$D_{KL}(P||Q) = \sum P(x) * log(P(x)) - \sum P(x) * log(Q(x)), \tag{3.29}$$

ou bien,

$$H(P,Q) = D_{KL} + S_p. \tag{3.30}$$

D'un point de vue d'optimisation, S_p est une constante car elle ne dépend pas du modèle $Q(x)$. Utiliser la divergence de Kullback-Leibler comme fonction objectif

dans un algorithme de classification binaire revient donc à minimiser l'entropie croisée. Un exemple est donné pour démontrer la mécanique de ce qui sera appliqué.

Example 3.2.1 *Si on suppose que P est une distribution Bernoulli(0.5) et que Q est une distribution Bernoulli($\hat{q}_i$), alors la divergence KL est*

$$D_{KL}(P||Q) = -\sum \left((0.5)\log(\hat{q}_i) + (0.5)\log(1-\hat{q}_i)\right) - S_p \tag{3.31}$$

qui aura le même optimum que la fonction Log-loss. Ainsi, un modèle dont les paramètres sont obtenus en minimisant la fonction

$$\begin{aligned} \ell^* = &-\sum_{i=1}^{n} \left(y_i \log(\hat{q}_i) + (1-y_i)\log(1-\hat{q}_i)\right) \\ &+ \lambda \left(\sum \left(y_i \log(\hat{q}_i) + (1-y_i)\log(1-\hat{q}_i)\right)\right) \end{aligned} \tag{3.32}$$

est un modèle dans lequel on choisit les paramètres de façon à minimiser la divergence entre la distribution Bernoulli obtenue et une distribution Bernoulli(0.5), c'est-à-dire en favorisant $\Pr(Y_i = 0) \approx \Pr(Y_i = 1)$.

Dans le contexte de ce travail, on cherche à obtenir des prédictions équitables au sens des groupes. Pour se faire, on veut minimiser la divergence KL entre une Bernoulli dont les paramètres sont estimés par Boosting et une Bernoulli dont les paramètres correspondent aux probabilités empiriques de survenance pour chaque sous-classe de l'attribut sensible. On suppose maintenant que P^F est une distribution Bernoulli(q^F) où

$$q^F = \frac{\sum_{i=1}^{n} \mathbb{I}\left(y_i = 1 \cap \text{Sexe}_i = F\right)}{\sum_{i=1}^{n} \mathbb{I}\left(\text{Sexe}_i = F\right)}, \tag{3.33}$$

que P^H est une distribution Bernoulli(q^H) où

$$q^H = \frac{\sum_{i=1}^{n} \mathbb{I}\left(y_i = 1 \cap \text{Sexe}_i = H\right)}{\sum_{i=1}^{n} \mathbb{I}\left(\text{Sexe}_i = H\right)} \tag{3.34}$$

et Q est une distribution Bernoulli($\hat{q}_i$). On a alors

$$\begin{aligned} D_{KL}(P^F||Q) &= -\sum \left(q^F \log(\hat{q}_i) + (1-q^F)\log(1-\hat{q}_i)\right) - S_{P^F} \\ D_{KL}(P^H||Q) &= -\sum \left(q^H \log(\hat{q}_i) + (1-q^H)\log(1-\hat{q}_i)\right) - S_{P^H} \end{aligned} \tag{3.35}$$

et la fonction à optimiser est

$$\begin{aligned} \ell^*_{Sexe} = &-\sum_{i=1}^{n} \left(y_i \log(\hat{q}_i) + (1-y_i)\log(1-\hat{q}_i)\right) \\ &+ \lambda_{Sexe}(-\sum_{i:\text{Sexe}_i=F} \left(q^F \log(\hat{q}_i) + (1-q^F)\log(1-\hat{q}_i)\right) \\ &- \sum_{i:\text{Sexe}_i=H} \left(q^H \log(\hat{q}_i) + (1-q^H)\log(1-\hat{q}_i)\right)). \end{aligned} \tag{3.36}$$

Les calculs du gradient et de la hessienne sont nécessaires pour la mise en oeuvre de l'algorithme et pour contrôler les prédictions des hommes et des femmes en les poussant à se rapprocher de la valeur empirique afin d'atténuer l'effet discriminatoire que la variable sensible pourrait avoir sur les prédictions. Le gradient mis en oeuvre dans l'algorithme de Boosting est donc

$$\begin{aligned} \frac{\partial \ell^*_{Sexe}}{\partial \beta_j} = \sum_{i}^{n} (\hat{q}_i - y_i) + \lambda_{Sexe}\Big(&- \sum_{i:\text{Sexe}_i=F} (q^F(1-\hat{q}_i) - (1-q^F)\hat{q}_i) \\ &- \sum_{i:\text{Sexe}_i=H} (q^H(1-\hat{q}_i) - (1-q^H)\hat{q}_i)\Big), \end{aligned} \tag{3.37}$$

et la hessienne est

$$\begin{aligned} \frac{\partial^2 \ell^*_{Sexe}}{\partial \beta_j \partial \beta_k} = \sum_{i}^{n} (\hat{q}_i(1-\hat{q}_i)) + \lambda_{Sexe}\Big(&- \sum_{i:\text{Sexe}_i=F} [q^F[(1-\hat{q}_i)^2 - (1-\hat{q}_i)] - (1-q^F)\hat{q}_i(1-\hat{q}_i)] \\ &- \sum_{i:\text{Sexe}_i=H} [q^H[(1-\hat{q}_i)^2 - (1-\hat{q}_i)] - (1-q^H)\hat{q}_i(1-\hat{q}_i)]\Big). \end{aligned} \tag{3.38}$$

Pour les variables sensibles qui ont plus que deux classes, comme l'âge, on a quatre groupes qui sont $A1$: 16 à 25, $A2$: 26 à 45, $A3$: 46 à 65 et $A4$: 66 *et plus*, la fonction d'optimisation va prendre en compte q^{A1}, q^{A2}, q^{A3} et q^{A4} qui sont les probabilités empiriques de survenance d'un sinistre pour chaque groupe d'âge. De même pour la variable $sexe \times \hat{a}ge$ qui a au total 8 classes, les quatre groupes d'âge pour les femmes ainsi que les quatre groupes d'âge pour les hommes.

Deux modèles équitables seront comparés également, le premier qui ne tient pas compte des proportions de chaque sous-classe de la variable sensible et le deuxième qui optimise la fonction objectif en prenant compte des proportions. Par exemple, pour la variable *sexe*, il y a 46.94% de femmes dans la base de données et 53.06% d'hommes. Cette différence est prise en compte dans la fonction objectif et le deuxième modèle est optimisé grâce à la fonction suivante

$$\begin{aligned}\ell^*_{Sexe} = &-\sum_{i=1}^{n}\left(y_i \log(\hat{q}_i) + (1-y_i)\log(1-\hat{q}_i)\right) \\ &+ \lambda_{Sexe}\Big(-\sum_{i:\text{Sexe}_i=F}\left(q^F \log(\hat{q}_i) + (1-q^F)\log(1-\hat{q}_i)\right) * w_F \\ &-\sum_{i:\text{Sexe}_i=H}\left(q^H \log(\hat{q}_i) + (1-q^H)\log(1-\hat{q}_i)\right)\Big),\end{aligned} \tag{3.39}$$

où w_F représente l'ajustement de la proportion des femmes dans les données afin de compenser la différence entre les deux sous-groupes

$$w_F = \frac{\sum_{i=1}^{n}\mathbb{I}(Sexe_i = H)}{\sum_{i=1}^{n}\mathbb{I}(Sexe_i = F)}. \tag{3.40}$$

Le gradient et la hessienne deviennent

$$\frac{\partial \ell^*_{Sexe}}{\partial \beta_j} = \sum_i^n (\hat{q}_i - y_i) + \lambda_{Sexe}\Big(- \sum_{i:\text{Sexe}_i=F} (q^F(1-\hat{q}_i) - (1-q^F)\hat{q}_i) * w_F \\ - \sum_{i:\text{Sexe}_i=H} (q^H(1-\hat{q}_i) - (1-q^H)\hat{q}_i)\Big), \tag{3.41}$$

$$\frac{\partial^2 \ell^*_{Sexe}}{\partial \beta_j \partial \beta_k} = \sum_i^n (\hat{q}_i(1-\hat{q}_i)) + \lambda_{Sexe}\Big(- \sum_{i:\text{Sexe}_i=F} [q^F[(1-\hat{q}_i)^2 - (1-\hat{q}_i)] - (1-q^F)\hat{q}_i(1-\hat{q}_i)] * w_F \\ - \sum_{i:\text{Sexe}_i=H} [q^H[(1-\hat{q}_i)^2 - (1-\hat{q}_i)] - (1-q^H)\hat{q}_i(1-\hat{q}_i)]\Big). \tag{3.42}$$

Il est à noter que lorsque $\lambda_{Sexe} = 0$, la fonction d'optimisation est équivalente à l'équation 3.17 du modèle de Boosting sans contrainte d'équité.

CHAPITRE IV

ANALYSE NUMÉRIQUE

Dans le présent chapitre, les résultats numériques seront analysés et discutés. Les valeurs du *p-rule* ainsi que les primes pures estimées pour quelques profils choisis sont aussi présentées pour les divers modèles.

La base de données utilisée dans les modèles est un échantillon de données en assurance automobile pour les couvertures de collision au Québec. Elle s'étale entre les années 2015 et 2021 et a été séparée aléatoirement avec une partition de 75% pour les données d'entrainement et de 25% pour les données test. Elle a été divisée de manière aléatoire afin d'éviter l'effet que la Covid-19 a eu sur les réclamations dans les années 2020 et 2021. Durant ces années, une diminution de la fréquence de réclamation ainsi qu'une augmentation de la sévérité a eu lieu, ce qui aurait pu affecter la performance des modèles.

4.1 Statistiques descriptives

Un total de 18 variables explicatives et de 473 866 observations sont utilisées pour l'entrainement du modèle, ainsi que 157 967 observations pour la validation du modèle. Quelques variables explicatives classiques de tarification sont présentées dans le tableau 4.1. La sélection de variables est abordée dans la section suivante.

Nom de la variable	Description	Min	Max	Moyenne
COST_INCURREDLOSS_M	Perte encourue de l'assuré	0	73 188.82	53.70
DRV_GENDER_A	Sexe de l'assuré	-	-	-
DRV_AGE_A	Âge de l'assuré	16 à 25	66 et plus	-
VEHICLEAGE	Âge du véhicule	0	98	5.51
COST_EARNED_EXPOSURE_M	Exposition	3.17e-08	1	0.32
DEDUCTIBLE_A	Déductible applicable	100	20 000	475.77
YEARSCLAIMFREE_LIABILITY	Nombre d'années sans réclamation	0	83	23.42

TABLE 4.1 – Quelques variables explicatives

Il y a un total de 468 477 observations dans la base de données qui n'ont pas de réclamation, ce qui fait en sorte qu'environ 1.14% des assurés observés ont subi une perte supérieure à 0. Comme mentionné dans le tableau 4.1, la réclamation moyenne pour tous les assurés est de 53.70$, tandis que la réclamation moyenne sachant que celle-ci est supérieure à 0 est de 4 722.33$. De plus, parmi toutes les observations, il y a un total de 64 469 polices distinctes. Une police distincte peut contenir plusieurs contrats différents et ils peuvent être renouveller à travers les années. Ceux-ci composent les 468 866 observations de la base de données. Dans le cadre de l'analyse présente, on supposera que les années et les contrats sont indépendantes.

Les variables *sexe* et *âge* sont choisies comme attributs sensibles pour mitiger la discrimination dans les algorithmes prédictifs. L'âge a été regroupé en 4 classes distinctes : les 16 à 25 ans, les 26 à 45 ans, les 46 à 65 ans et les 66 ans et plus. Le nombre d'observations ainsi que les pertes moyennes liées à chaque sous-groupe de ces deux variables sensibles sont présentés dans les tableaux 4.2, 4.3 et 4.4. Les figures 4.1 et 4.2 permettent de visualiser les résultats des tableaux mentionnés précédemment. On voit bien qu'en général, pour les hommes et pour les femmes, la sous-classe de 16 à 25 ans est celle qui constitue les pertes moyennes les plus élevées. Ceci diminue drastiquement pour la sous-classe des âges 26 à 45 ans comme on peut le voir à la Figure 4.1. On peut donc s'imaginer qu'il y a beaucoup moins de réclamations dans cette tranche d'âge. Pour les pertes supérieures à 0, à la Figure 4.2, on voit une diminution également à travers toutes les tranches d'âges pour finir avec une légère augmentation pour les femmes de 66 ans et plus. Cette tendance est aussi apparente à la Figure 4.1.

Sexe	Nombre d'obs.	Perte moy. ($)	Perte moy. supérieure à 0 ($)
Femme	222 457	51.68	4 397.78
Homme	251 409	55.50	5 028.06

TABLE 4.2 – Statistiques descriptives propre à la variable *sexe*.

Âge	Nombre d'obs.	Perte moy. ($)	Perte moy. supérieure à 0 ($)
16 à 25	29 330	113.76	6 110.72
26 à 45	182 663	59.38	4 955.13
46 à 65	194 106	40.42	4 320.68
66 et plus	67 767	50.45	4 080.00

TABLE 4.3 – Statistiques descriptives propre à la variable *âge*.

Sexe-âge	Nombre d'obs.	Perte moy. ($)	Perte moy. supérieure à 0 ($)
16 à 25 F	16 093	98.93	5 527.80
26 à 45 F	93 785	55.11	4 639.55
46 à 65 F	86 734	38.29	3 893.82
66 et plus F	25 845	54.71	3 938.44
16 à 25 H	13 237	131.79	6 761.43
26 à 45 H	88 878	63.89	5 282.17
46 à 65 H	107 372	42.14	4 698.79
66 et plus H	41 922	47.83	4 186.09

TABLE 4.4 – Statistiques descriptives propre aux variables *sexe* et *âge* combinées.

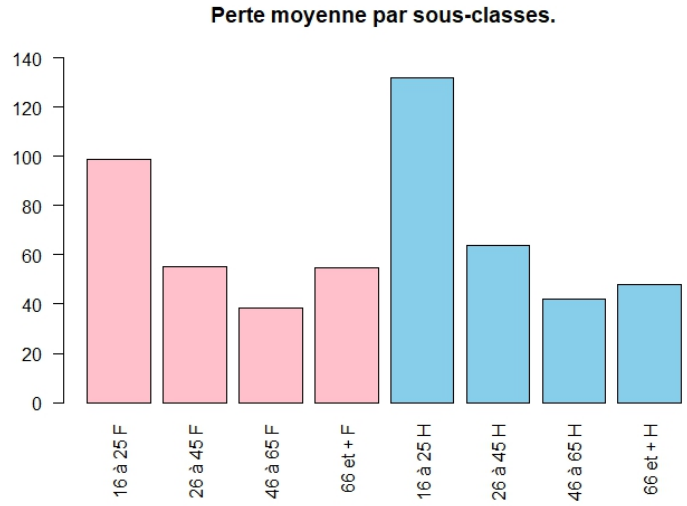


FIGURE 4.1 – Perte moyenne selon les différentes sous-classes sur lesquelles l'équité des modèles est contrôlée.

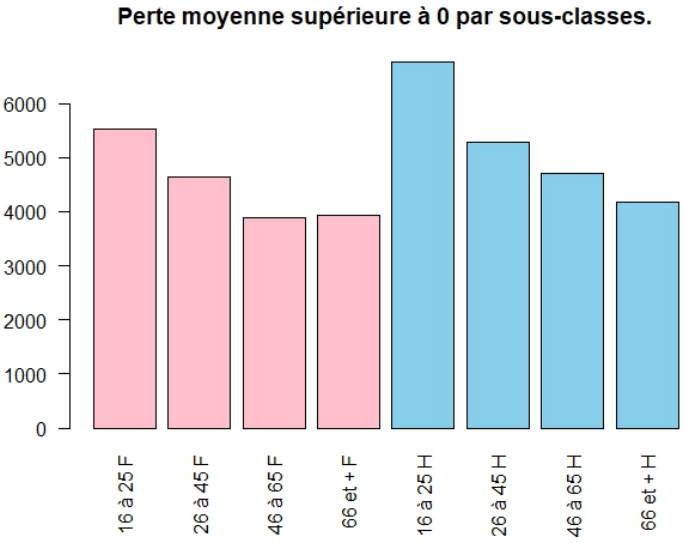


FIGURE 4.2 – Perte moyenne supérieure à 0 selon les différentes sous-classes sur lesquelles l'équité des modèles est contrôlée.

4.2 Aspect computationnel

Les langages de programmation utilisés sont R et Python. R a principalement été utilisé pour les MLG et les modèles de Boosting Bernoulli et Gamma (modèle survenance-sévérité). La manipulation et l'analyse de données ont été faites à l'aide de la librairie *dplyr*, la sélection de variables et le réglage des hyperparamètres ont été faits à l'aide des librairies *tweedie* et *HDtweedie*. Les différents modèles MLG et Boosting Bernoulli ont été entrainés grâce aux librairies *caret*, *stats* et *xgboost*. Le temps de calculs pour chaque modèle était de quelques minutes. Le calcul des hyperparamètres optimaux, cependant, était un peu plus long et durait environ une vingtaine de minutes. Le langage de programmation Python a surtout été utilisé pour les modèles de Boosting Tweedie avec et sans contrainte d'équité. La librairie *pandas* a été utilisée pour l'analyse et la manipulation de données. Le réglage des hyperparamètres et les modèles ont été faits à l'aide des librairies *xgboost*, *numpy*, *sklearn* et *tweedie*. Finalement, la librairie *numdifftools* a été utilisée pour la différentiation numérique de la divergence de Kullback-Leibler. Le temps de calcul pour le modèle Boosting Tweedie sans contrainte d'équité prenait quelques minutes et celui avec la fonction objectif personnalisée ne convergeait pas. Une approche avec la modélisation de la survenance et de la sévérité séparément a donc été réalisée et sera détaillée ultérieurement.

4.3 MLG et Boosting Tweedie

Le modèle linéaire généralisé Tweedie est comparé avec les deux modèles de Boosting Tweedie, avec et sans contrainte d'équité. La performance du modèle pour les variables sensibles choisies est calculée avec le ratio de la divergence KL de l'équation 3.25 afin d'évaluer la différence entre les modèles équitables et non équitables. Pour une équité parfaite, ce ratio devrait être égal à 1.

4.3.1 Sélection de variables

La sélection de variables pour le MLG Tweedie a été effectuée à l'aide d'une pénalisation Lasso. Le paramètre ω (*shrinkage parameter*) a été choisi par validation croisée, et celui qui correspondait à l'erreur moyenne minimale par validation croisée (*mean cross-validated error*) a été retenue. Ceci équivaut à un $\omega = 0.2490605$ pour une erreur moyenne de 61.70. La pénalisation Lasso est donnée par

$$argmin_{\beta_j}\Big(\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \omega\sum_{j=1}^{p}|\beta_j|\Big) \tag{4.1}$$

et la valeur des paramètres les moins importants est fixée à 0. Dans la figure 4.3, on voit le rétrécissement des β_j selon le logarithme des différents ω. Plus celui-ci augmente, plus la pénalisation est sévère. Les lignes représentent les valeurs des coefficients β_j qui atteignent tous une valeur nulle. La ligne noire verticale indique le modèle qui a la plus petite erreur moyenne et les 18 variables conservées parmi les 34 pour l'entrainement.

Les mêmes variables explicatives sont utilisées pour tous les modèles afin de permettre d'analyser uniquement l'impact du contrôle de l'équité.

4.3.2 MLG Tweedie

Le paramètre ρ optimal qui a été calculé avec les variables choisies pour le premier modèle de tarification automobile est de 1.414082. Le choix de ρ est fait en maximisant la vraisemblance Tweedie comme présenté dans la figure 4.4. Un paramètre ρ entre 1 et 2 permet d'obtenir une distribution Poisson-Gamma lorsqu'il y a une masse à 0, ce qui est le cas dans les données en assurance. En effet, les figures 4.5a et 4.5b permettent d'illustrer ce phénomène. La figure 4.5a est une représentation

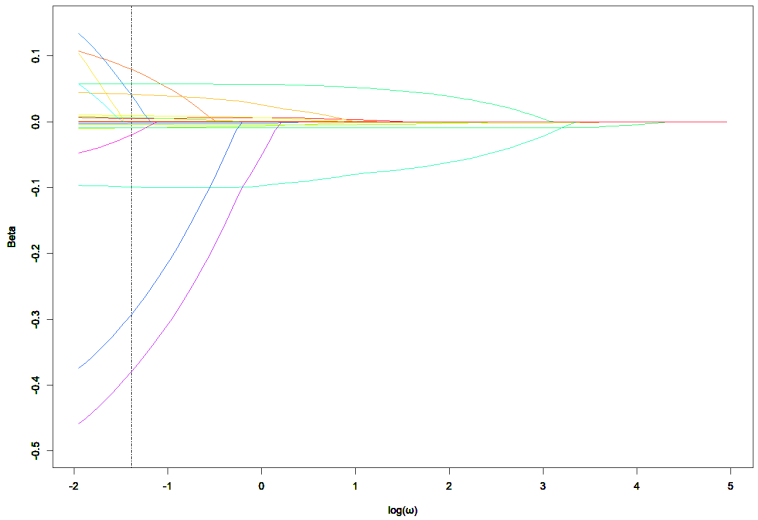


FIGURE 4.3 – Pénalisation Lasso

graphique des pertes encourues en incluant la masse à 0 présente dans les données, alors que la figure 4.5b représente les pertes encourues en excluant cette masse. Il est difficile de visualiser les pertes encourues supérieures à 0 si on ne l'exclut pas.

Le paramètre ϕ est estimé par l'algorithme du MLG Tweedie et ceci donne un $\hat{\phi} = 3450.434$. La variance du modèle est donc

$$Var(Y_i) = 3450.434\mathbb{E}(Y_i)^{1.414082}. \tag{4.2}$$

Les mêmes valeurs de ϕ et ρ sont utilisées pour l'algorithme de Boosting Tweedie. Encore une fois, ceci permet d'analyser uniquement l'impact du contrôle de l'équité. Le logarithme de la variable d'exposition *(COST_EARNED_EXPOSURE_M)* est utilisé comme un *offset* dans le MLG Tweedie et on utilise une fonction de lien logarithmique.

Deux métriques d'évaluation de la performance du modèle sont calculées sur les

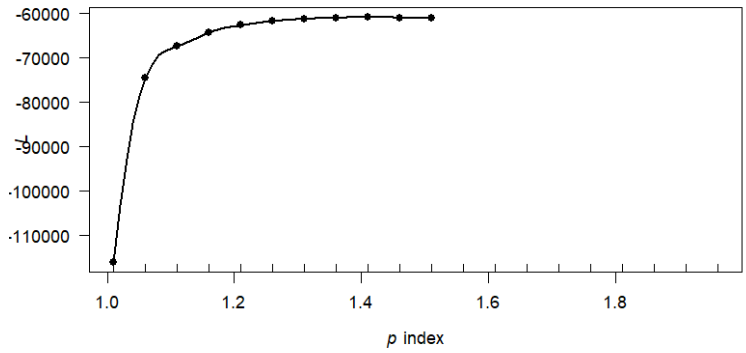
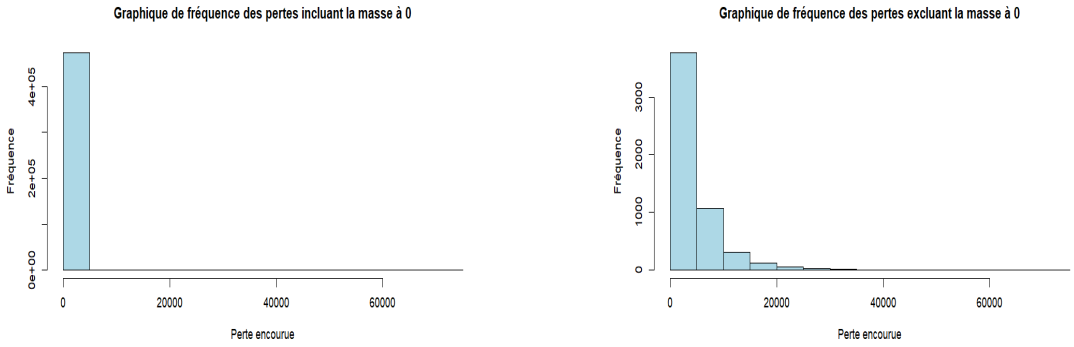


FIGURE 4.4 – Rho qui minimise le log-loss



(a) (b)

FIGURE 4.5 – Graphique de fréquence Tweedie selon les données.

prédictions issues des données de validation. La première est la racine de l'erreur quadratique moyenne ($RMSE$) telle que présentée à l'équation 4.3 et la seconde, qui est présentée à l'équation 4.4, est l'erreur absolue moyenne (MAE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{4.3}$$

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N}. \tag{4.4}$$

On a N qui représente le nombre total d'observations sur lesquelles ces métriques sont calculées. Pour le MLG Tweedie, le $RMSE = 817.60$ et le $MAE = 114.81$. La somme des prédictions de ce modèle est égale à 9 378 994 alors que la somme des vrais pertes est égale à 9 182 118. Cette différence est dûe au fait que le lien logarithmique, qui n'est pas le lien canonique de la famille Tweedie, a été utilisé pour l'entrainement de ce modèle. Dans le cas où le lien canonique n'est pas utilisé, un facteur de rebalancement doit être considéré afin d'égaliser la somme des prédictions avec la somme des vrais pertes. Ce facteur devrait aussi être appliqué sur chaque prédictions individuellement. Ce dernier est égale à 1.021441 dans ce contexte-ci. Dans le cadre de cette recherche un facteur de rebalancement n'a pas été considéré et les résultats du $RMSE$ et du MAE ont été calculés avec les préditions sans en tenir compte.

4.3.3 Boosting Tweedie

Les variables explicatives pour le modèle de Boosting Tweedie sont les mêmes que celles présentées dans la sous-section précédente. Un modèle de tarification sans contrainte d'équité est fait et les hyperparamètres optimaux ajustés sont présentés au tableau 4.5. Cinq hyperparamètres ont été testés afin d'obtenir leur

valeur optimale pour l'entrainement du modèle. Parmi ceux-ci on retrouve *learning_ rate (eta)* qui prévient le sur-ajustement du modèle, *max_ depth* qui indique la profondeur maximale de l'arbre de décision, *min_ child_ weight* qui indique le nombre minimal d'instance dans chaque noeud, *subsample*, la proportion des données d'entrainement sous-échantillonées qui a lieu à chaque itération et, finalement, *colsample_ bytree*, la proportion des variables de la base de données sélectionnées aléatoirement qui sera utilisée pour l'entrainement de chaque arbre de décision.

Hyperparamètre	**Valeur retenue**
learning_ rate (eta)	0.05
max_ depth	3
min_ child_ weight	5
subsample	0.5
colsample_ bytree	1

TABLE 4.5 – Hyperparamètres du modèle de Boosting Tweedie.

Le $RMSE$ de ce modèle est égal à 771.06 et le MAE est égal à 10.15. Il y a une nette amélioration dans les prédictions faites par le modèle de Boosting sans contrainte d'équité comparativement au MLG Tweedie.

Pour le Boosting avec contrainte d'équité, la mise en oeuvre de l'équation 3.26 dans l'algorithme nécessite du temps et des compétences en programmation qui sont hors de l'expertise de ce travail de recherche. La fonction de perte suivante a donc été mise en oeuvre

$$L_{D_{KL}} = -\sum_i y_i \frac{e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}*(1-\rho)}}{1-\rho} + \frac{e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}*(2-\rho)}}{2-\rho} + \lambda * D_{KL}(F^*(\boldsymbol{x}), 1). \tag{4.5}$$

Le calcul de la divergence KL de chaque sous-groupe distinct, tel que présenté aux

équations 3.23 et 3.24, est combiné en une seule divergence KL qui est calculée sur l'ensemble des données. Le calcul de la parité démographique sur la divergence de Kullback-Leibler (équation 3.25) est ensuite calculé pour chaque sous-groupe afin d'évaluer la performance du modèle sur les variables sensibles avec l'ajout de la contrainte d'équité.

L'algorithme n'a pas convergé avec succès car l'exécution du code était très longue. En effet, l'ajout du calcul de la divergence KL entre la répartition Tweedie et la densité uniforme à la fonction d'optimisation du modèle rendait l'algorithme très lourd. Subséquemment, les résultats ne sont pas concluants pour ces modèles.

4.4 MLG et Boosting Bernoulli

La même base de données ainsi que les mêmes variables explicatives, utilisées pour le MLG Tweedie présenté plus haut, sont utilisées pour ce modèle. L'effet des variables sensibles est contrôlé dans le modèle de Boosting Bernoulli. Effectivement, la sévérité d'un accident ne dépend pas nécessairement des attributs d'un assuré et la discrimination n'est pas évaluée sur le MLG de sévérité. Toutefois, la probabilité de survenance ou pas d'un sinistre pourrait être affectée négativement par les variables sensibles d'un assuré, comme le sexe ou l'âge.

4.4.1 MLG de survenance et de sévérité

Dans le MLG Bernoulli, l'équité est quantifiée à l'aide du *p-rule* de l'équation 2.1. Si ce ratio est proche de 1, cela indique que l'algorithme est équitable. Cependant, au lieu de comparer les probabilités de survenance entre les sous-groupes des attributs sensibles, la probabilité empirique de survenance du sous-groupe A est comparée avec la moyenne des probabilités prédites pour le même sous-groupe. En prenant comme variable sensible le sexe, la probabilité empirique de survenance

d'un sinistre, pour les femmes et pour les hommes, est calculée par les équations 3.33 et 3.34. Le *p-rule* est calculé pour le MLG Bernoulli et les modèles de Boosting Bernoulli sans et avec la divergence KL comme contrainte d'équité.

Pour le MLG qui modélise la sévérité, une distribution de famille Gamma est choisie. Les données pour l'entrainement sont filtrées afin de prendre ceux qui ont une sévérité strictement supérieure à 0. Ceci nous donne 6 597 observations au total. Les variables explicatives utilisées ne sont pas les mêmes que les autres modèles. En partant des mêmes variables initiales, elles ont été choisies par un test ANOVA de type II (test-F) et le modèle résultant donne le AIC le plus bas parmi tous les modèles possibles qui est de 124 677. Au total, 9 variables explicatives sont conservées. De plus, un lien logarithmique est utilisé dans cet algorithme. La prime pure finale, PP, est calculée en multipliant la prédiction de la probabilité de survenance de l'assuré i du MLG Bernoulli avec la prédiction de la sévérité, L, du même assuré

$$\mathbb{E}[PP] = \mathbb{E}[Z] \times \mathbb{E}[L]. \tag{4.6}$$

Les métriques de performance de ces modèles sont également calculées et le $RMSE$ entre la vraie perte encourue et la prime pure calculée est égal à 817.39 et le MAE est égal à 110.42. Il y a une légère amélioration entre la performance du MLG Tweedie, mais ce n'est pas un changement significatif.

4.4.2 Boosting Bernoulli

Le premier algorithme de Boosting est ensuite mis en oeuvre en ajoutant la divergence KL comme métrique d'équité pour viser spécifiquement la variable sensible *sexe* tel que montré à l'équation 3.39. Le poids w_F nécessaire pour l'ajustement

des proportions de chaque sous-classe est ensuite pris en compte afin de tester si la rectification des proportions est avantageuse. Pour la variable *sexe*, le nombre d'observations appartenant à chaque sous-classe est affiché dans la table 4.2. Ceci donne une proportion d'environ 53.05% d'hommes et 46.95% de femmes dans la base de données. Dans ce cas-ci, $w_F = 1.130147$. De plus, les hyperparamètres optimaux ajustés pour ce modèle sont présentés au tableau 4.6.

Hyperparamètre	**Valeur retenue**
eta	0.3
max_ depth	2
min_ child_ weight	1
subsample	1
colsample_ bytree	0.8
nrounds	100

TABLE 4.6 – Hyperparamètres du modèle Boosting Bernoulli.

Ce sont les mêmes hyperparamètres qui sont utilisés pour les algorithmes qui contrôlent les variables *âge* et *sexe*×*âge*, ainsi que ceux qui ne prennent pas en compte les poids pour l'ajustement des proportions de chaque sous-classe. Une pondération, λ, est aussi donnée à la contrainte d'équité dans l'équation d'optimisation pour le contrôle du compromis performance-équité. Plusieurs valeurs de λ se situant entre 0 et 1 avec des bonds de 0.1 ont été expérimentées. Les métriques de performance selon chaque λ pour ce premier modèle sont présentées dans le tableau 4.7. Le Boosting sans contrainte d'équité à un $RMSE = 823.88$ et un $MAE = 110.85$.

Pour le modèle qui contrôle l'attribut sensible *âge*, les paramètres λ expérimentés se situent entre 0 et 1.5. Les proportions de chaque sous-classe de la variable *âge* sont notés au tableau 4.3. Les poids sont calculés par rapport à la troisième

λ	$RMSE$	MAE
0.1	824.89	111.21
0.2	827.27	111.47
0.3	829.31	111.77
0.4	832.38	112.06
0.5	829.82	112.06
0.6	831.49	112.25
0.7	835.19	112.53
0.8	834.86	112.57
0.9	839.20	112.82
1	835.15	112.72

TABLE 4.7 – Métrique de performance pour le modèle Boosting de survenance avec fonction objectif qui contrôle la variable *sexe* selon les différentes valeurs de λ. La proportion des sous-classes est prise en compte.

sous-classe puisqu'elle est la plus nombreuse. On a donc $w_{A1} = 6.618002$, $w_{A2} = 1.062645$ et $w_{A4} = 2.864314$. Les métriques de performance selon chaque valeur de λ sont présentées dans le tableau 4.8.

Finalement, pour le modèle qui contrôle la variable intersectionnelle, *sexe*×*âge*, le tableau 4.4 indique que la sous-classe la plus nombreuse est celle des hommes âgés entre 46 et 65 ans. Les poids sont donc calculés par rapport à cette classe et on obtient les valeurs présentées au tableau 4.9.

Pour ce qui en est des métriques de performance selon les différents λ, les valeurs sont présentés au tableau 4.10.

Pour les trois variables sensibles présentées plus haut, on peut voir une tendance à la hausse des métriques de performance plus qu'on augmente le poids λ qu'on

λ	$RMSE$	MAE
0.1	836.53	112.00
0.2	837.86	112.43
0.3	837.04	112.68
0.4	838.10	112.90
0.5	837.03	112.98
0.6	844.87	113.46
0.7	842.21	113.46
0.8	841.55	113.51
0.9	845.42	113.73
1	843.94	113.79
1.1	844.82	113.82
1.2	846.48	113.98
1.3	846.68	114.05
1.4	846.46	114.10
1.5	846.96	114.07

TABLE 4.8 – Métrique de performance pour le modèle Boosting de survenance avec fonction objectif qui contrôle la variable *âge* selon les différentes valeurs de λ. La proportion des sous-classes est prise en compte.

donne à l'équité. Cette observation va avec ce qui a été discuté au chapitre 2 concernant le compromis précision-équité du modèle. Pour le sexe, l'âge et l'intersection entre le sexe et l'âge, on observe le même phénomène pour le $RMSE$ et le MAE. On s'aperçoit que le $RMSE$ est beaucoup plus affecté que le MAE, mais ceci est dû au fait que le MAE est une métrique beaucoup plus robuste puisqu'elle n'élève pas les erreurs au carré comme le fait le $RMSE$. En effet, en la présence de valeurs extrêmes, il est préférable d'utiliser le MAE comme métrique

Poids	**Valeur**
w_{A1-F}	6.671969
w_{A2-F}	1.144874
w_{A3-F}	1.237946
w_{A4-F}	4.154459
w_{A1-H}	8.111506
w_{A2-H}	1.208083
w_{A4-H}	2.561233

TABLE 4.9 – Poids pour les sous-classes de la variable *sexe*×*âge*.

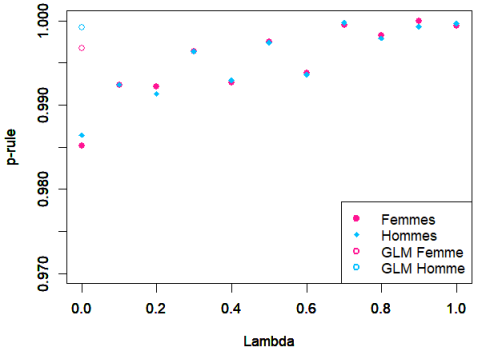
λ	$RMSE$	MAE
0.1	832.29	111.85
0.2	835.35	112.30
0.3	837.29	112.74
0.4	837.64	112.85
0.5	837.51	113.08
0.6	837.38	113.23
0.7	841.26	113.51
0.8	842.02	113.55
0.9	843.06	113.71
1	842.15	113.80

TABLE 4.10 – Métrique de performance pour le modèle Boosting de survenance avec fonction objectif qui contrôle la variable *sexe*×*âge* selon les différentes valeurs de λ. La proportion des sous-classes est prise en compte.

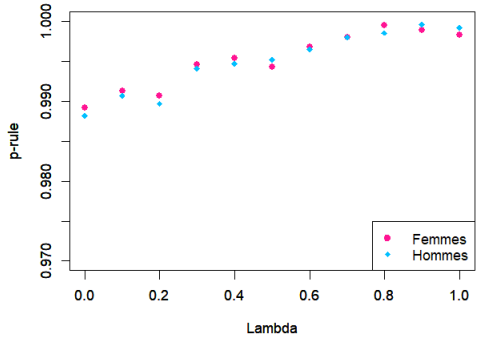
de performance puisqu'elle est, en général, plus robuste aux valeurs aberrantes.

Le *p-rule* selon les 14 différentes sous-classes abordées précédemment est modélisé dans les figures 4.6, 4.7, 4.8 et 4.9. Plus λ augmente, plus on s'attend à voir une amélioration de la métrique *p-rule* et donc un ratio proche de 1. En prenant le premier graphique 4.6, le *p-rule* sur le groupe des hommes et des femmes est calculé sur le modèle MLG Bernoulli ainsi que sur le modèle Boosting Bernoulli avec différents λ. Le *p-rule* du MLG Bernoulli est déjà très proche de 1, la valeur souhaitée, puisque les prédictions dans un MLG sont équitables au sens des groupes et les probabilités prédites se rapprochent des probabilités empiriques de chaque groupe respectivement. Par contre, l'algorithme de Boosting ne respecte pas cette contrainte et donc le même résultat est visé dans le deuxième modèle qui est le Boosting de classification. On peut voir l'augmentation du *p-rule* avec l'augmentation du λ jusqu'à même dépasser celle du MLG Bernoulli. De plus, en comparant le graphique 4.6a au graphique 4.6b, on constate que l'ajout des poids pour équilibrer les proportions des femmes par rapport aux hommes dans les données fait réellement une différence. En effet, avec l'ajout des poids, le *p-rule* converge plus rapidement vers 1. Ceci est encore mieux perçu dans les graphiques suivants, où la variable sensible contient plus que deux sous-classes et où ceux-ci ont plus de disparités dans leurs proportions.

En observant les figures 4.7a et 4.7b, l'effet des poids peut être mieux perçu. Encore une fois, le *p-rule* du MLG Bernoulli est proche de 1 pour les 4 sous-classes de la variable sensible âge. Pour le modèle de Boosting, la métrique d'équité converge vers 1 beaucoup plus rapidement lorsque les poids sont présents dans la contrainte d'équité. De plus, les sous-classes les plus affectées par cet ajout sont celles qui sont les moins nombreuses. En effet, comme présenté au tableau 4.3, la sous-classe 16 à 25 ans (ligne orange dans la figure 4.7) contient 29 330 observations et la sous-classe 66 et plus (ligne bleu dans la figure 4.7) en contient 67 767. Le sous-groupe âgé entre 16 et 25 ans est donc celui qui est le plus avantagé par l'équilibre des



(a) Avec les poids.



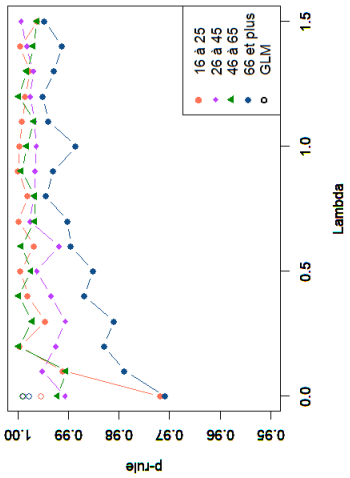
(b) Sans les poids.

FIGURE 4.6 – Le *p-rule* pour la variable sensible sexe.

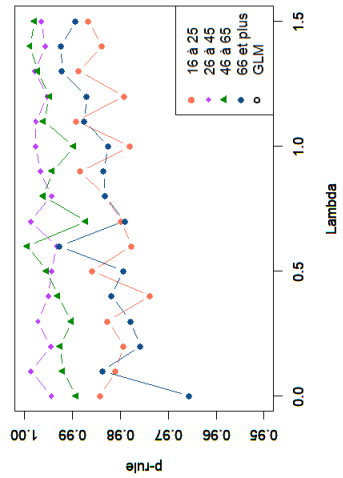
poids. On peut voir que le *p-rule* converge vers 1 beaucoup plus rapidement que sans la présence de poids. Le deuxième groupe qui en bénéficie et le sous-groupe de ceux âgés de 66 ans et plus. Le modèle contrôle assez bien les 2 sous-groupes restants puisqu'ils sont quand même assez nombreux et l'algorithme réussit à les représenter de manière adéquate.

Pour l'intersection des attributs sensibles sexe et âge, les mêmes conclusions peuvent en être tirées. Le graphique pour les femmes et celui des hommes ont été tracées séparément pour la clarté de la visualisation des résultats. Le tableau 4.4 indique que les deux sous-classes les moins nombreuses pour les femmes sont celles appartenant au groupe d'âge 16 à 25 ans et 66 ans et plus, avec respectivement 16 093 et 25 845 observations. Pour les hommes c'est le groupe d'âge 16 à 25 ans qui est le moins nombreux avec 13 237 observations. Il est donc attendu que ce soient ces sous-groupes qui bénéficient le plus par l'ajout d'un poids, ce qui est exactement ce qui est observé dans les figures 4.8 et 4.9. Également, avec la présence des poids dans l'algorithme, les sous-groupes semblent tous converger vers la même valeur en étant stable, contrairement aux algorithmes sans poids, où chaque sous-groupe converge à une valeur différente.

Les primes pures sont calculées dans chaque modèle, le MLG Bernoulli, le Boosting sans contrainte d'équité et les divers Boosting avec les diverses contraintes d'équité, avec et sans poids, pour chaque variable sensible considérée dans ce travail de recherche. Au total, 16 profils sont choisis et leurs primes pures sont présentées aux tableaux 4.12 et 4.13. Les caractéristiques des 16 profils sont présentées au tableau 4.11. En observant cet échantillon de données, on peut s'apercevoir que les primes calculées avec les modèles de Boosting qui ne prennent pas en compte les proportions de chaque sous-classe (Figure 4.13) semblent être assez comparable, mais significativement différente des primes estimées par le MLG et le Boosting avec $\lambda = 0$. Alors qu'en prenant en compte les proportions (Figure

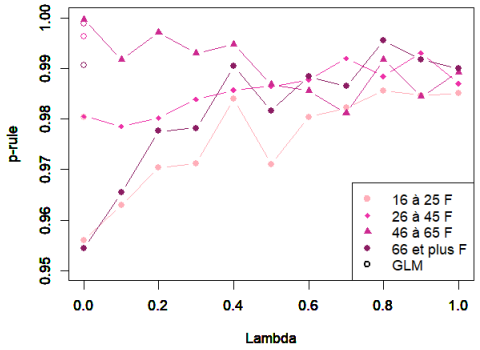


(a) Avec les poids.

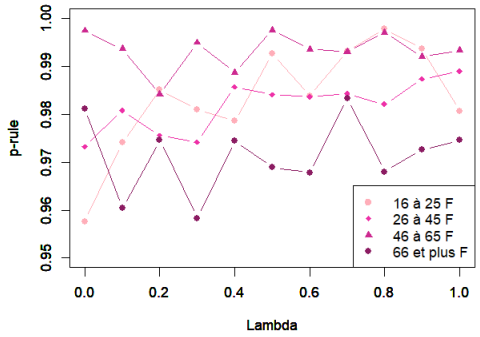


(b) Sans les poids.

FIGURE 4.7 – Le *p-rule* pour la variable sensible âge.

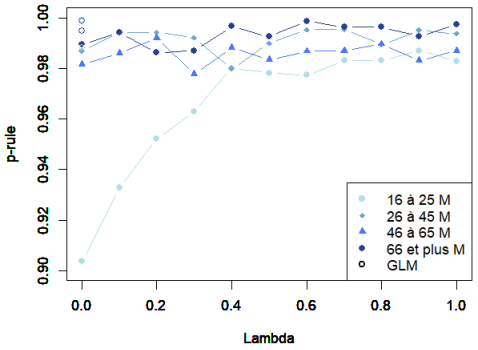


(a) Avec les poids.

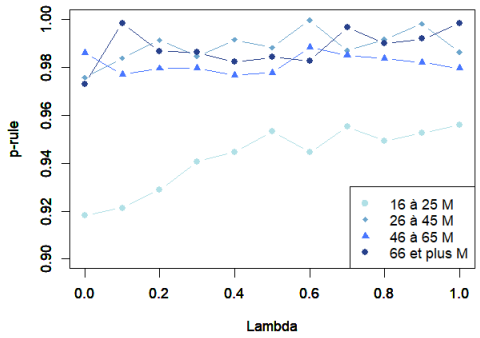


(b) Sans les poids.

FIGURE 4.8 – Le *p-rule* pour l'intersection des variables sensibles âge et sexe (Femme seulement).



(a) Avec les poids.



(b) Sans les poids.

FIGURE 4.9 – Le *p-rule* pour l'intersection des variables sensibles âge et sexe (Homme seulement).

4.12), les primes estimées par les modèles de Boosting se rapprochent un peu plus de ceux du MLG et ceci nous permet d'établir une équité au sens des groupes. En effet, comme mentionné précédemment les modèles linéaires généralisés respectent cette définition d'équité.

Le λ optimal est choisi selon les figures 4.6, 4.7, 4.8 et 4.9. Pour la variable sensible sexe, le λ qui équivaut au *p-rule* le plus élevé pour les femmes et pour les hommes simultanément est retenu. Celui-ci est égal à 0.7 pour l'algorithme avec poids et à 0.9 pour celui sans les poids.

L'ajout d'une contrainte d'équité fait bien une différence dans la prédiction des probabilités de survenance. En effet, les mêmes prédictions de la sévérité faite par le MLG Gamma sont utilisées dans le calcul de toutes les primes pures, ce qui est affecté par la contrainte d'équité est donc la prédiction de la probabilité de survenance d'un sinistre pour chaque profil distinct. Ceci est fait de sorte à minimiser le *p-rule*, ce qui rend ces primes plus équitables selon la contrainte utilisée.

Le choix des λ optimaux équivaut également aux moins bonnes valeurs des $RMSE$ et MAE aux tableaux 4.7, 4.8 et 4.10. Il y a donc bel et bien un compromis équité-précision dans ces modèles. Par exemple, pour un $\lambda_{Sexe} = 0.7$, on retrouve un $RMSE = 835.19$, ce qui équivaut au neuvième modèle sur dix en matière de performance, et un $MAE = 112.53$, ce qui équivaut au septième modèle sur dix en matière de performance selon cette métrique. De plus, les λ optimaux sont les mêmes pour toutes les sous-classes de la variable sensible prise en compte. Ça pourrait être intéressant de prendre en compte différents λ pour différentes sous-classes du même attribut sensible et d'analyser les résultats qui en découlent. Pour la variable *sexe*, par exemple, l'ajout d'une pondération différente pour les hommes et les femmes dans la fonction d'optimisation, λ_{Homme} et λ_{Femme}, au lieu

d'un seul λ_{Sexe} commun pourrait donner des conclusions intéressantes.

En synthèse, cinq modèles distincts ont été mis en oeuvre et étudiés. Premièrement, le modèle standard, qui est le modèle linéaire généralisé Tweedie avec paramètres $\hat{\rho} = 1.414082$ et $\hat{\phi} = 3450.434$, a obtenu un $RMSE$=817.60 et un MAE=114.81. Comparativement à celui-ci, le modèle de Boosting Tweedie, qui a été mis en oeuvre avec le même paramètres $\hat{\rho}$ et $\hat{\phi}$, ainsi que les mêmes variables explicatives, a obtenu un $RMSE$=771.06 et un MAE=10.15 ce qui démontre une nette amélioration dans la performance. Toutefois, l'impact de l'ajout de la divergence de Kullback-Leibler en tant que contrainte d'équité a réellement eu un impact dans les modèles de survenance-sévérité. Le MLG Bernoulli servait à prédire la probabilité de survenance d'un sinistre et le MLG Gamma prédisait la sévérité de ce dernier. Les métriques de performance pour ces modèles donnaient un $RMSE$=817.39 et un MAE=110.42, ce qui est très proche des valeurs du modèle standard. Quant à la performance du modèle de Boosting Bernoulli, on a obtenu un $RMSE$=823.88 et un MAE=110.85, ce qui n'indique pas nécessairement une réelle amélioration comparativement au MLG Bernoulli. Par contre, l'équité du modèle est améliorée et le *p-rule* se raproche de 1 lorsqu'on augmente la valeur de λ. Il est ensuite possible d'observer que, plus un poids important est assigné à la contrainte d'équité, plus la performance du modèle décline, ce qui confirme les suppositions de base concernant le compromis équité-performance. Finalement, lorsque la proportion des classes est ajustée durant l'entrainement des modèles Boosting Bernoulli, le *p-rule* s'améliore et converge vers 1 plus rapidement, mais les métriques de performance semblent toujours augmenter lorsque λ croît.

Identifiant de profil	Sexe	Âge	Classe du conducteur	Déductible	Côte de crédit harmonisée	Territoire	Code urbain/rural
15	F	16 à 25	3	500	J	872	U
8 431	F	16 à 25	20	1000	R	876	R
339	F	26 à 45	1	500	L	879	R
91 400	F	26 à 45	2	2500	T	870	U
2 875	F	46 à 65	2	1000	K	874	R
19 042	F	46 à 65	3	1000	I	873	U
60 862	F	66 et plus	752	500	N	887	R
8 584	F	66 et plus	77	500	I	877	R
2 440	H	16 à 25	77	1000	P	870	U
11 135	H	16 à 25	152	500	J	873	U
2 007	H	26 à 45	252	500	G	870	U
128 429	H	26 à 45	452	250	Q	874	R
111 787	H	46 à 65	552	250	O	870	R
5	H	46 à 65	2	500	K	872	U
544	H	66 et plus	0	500	F	875	U
62 206	H	66 et plus	0	500	P	887	R

TABLE 4.11 – Caractéristiques des 16 profils choisis.

Identifiant de profil	**MLG**	**Boosting** $\lambda = 0$	**Boosting avec** $\lambda_{Sexe} = 0.7$	**Boosting avec** $\lambda_{Age} = 1.2$	**Boosting avec** $\lambda_{Sexe \times age} = 0.8$
15	52.02$	71.89$	74.89$	105.91$	93.12$
8 431	171.76$	137.25$	144.91$	175.50$	161.28$
339	45.31$	53.96$	50.31$	55.61$	57.57$
91 400	159.03$	219.57$	215.20$	174.58$	184.48$
2 875	40.05$	39.30$	42.60$	45.34$	41.77$
19 042	26.73$	43.42$	43.89$	37.34$	41.24$
60 862	103.49$	97.51$	147.60$	157.14$	148.64$
8 584	32.42$	36.03$	44.81$	48.84$	49.71$
2 440	59.24$	33.52$	59.36$	136.85$	132.40$
11 135	28.22$	47.42$	50.06$	73.27$	72.71$
2 007	32.63$	85.62$	94.50$	73.98$	76.69$
128 429	11.91$	58.48$	55.49$	63.47$	70.17$
111 787	139.71$	328.96$	264.25$	198.70$	208.29$
5	38.49$	43.85$	53.34$	49.08$	48.30$
544	70.04$	87.37$	72.89$	76.13$	72.77$
62 206	35.07$	37.94$	40.05$	49.22$	46.45$

TABLE 4.12 – Primes pures calculées dans les différents modèles en prenant en compte les proportions entre les sous-classes. Les Boosting avec les λ optimaux sont utilisés pour chaque variable sensible.

Identifiant de profil	**MLG**	**Boosting** $\lambda = 0$	**Boosting avec** $\lambda_{Sexe} = 0.9$	**Boosting avec** $\lambda_{Age} = 1.3$	**Boosting avec** $\lambda_{Sexe \times age} = 0.7$
15	52.02$	71.89$	75.51$	90.59$	73.27$
8 431	171.76$	137.25$	50.95$	43.54$	47.89$
339	45.31$	53.96$	49.52$	54.65$	49.19$
91 400	159.03$	219.57$	25.31$	29.49$	21.53$
2 875	40.05$	39.30$	37.23$	43.11$	39.46$
19 042	26.73$	43.42$	70.29$	62.77$	62.10$
60 862	103.49$	97.51$	37.21$	39.73$	40.42$
8 584	32.42$	36.03$	26.29$	23.58$	21.24$
2 440	59.24$	33.52$	33.61$	36.73$	34.82$
11 135	28.22$	47.42$	114.35$	87.51$	99.18$
2 007	32.63$	85.62$	127.42$	120.73$	120.34$
128 429	11.91$	58.48$	24.42$	22.68$	18.54$
111 787	139.71$	328.96$	67.19$	67.92$	71.94$
5	38.49$	43.85$	48.78$	47.15$	46.26$
544	70.04$	87.37$	82.40$	83.14$	84.09$
62 206	35.07$	37.94$	60.77$	51.71$	61.47$

TABLE 4.13 – Primes pures calculées dans les différents modèles sans prendre en compte les proportions entre les sous-classes.

CONCLUSION

En conclusion, la mitigation de la discrimination dans les algorithmes d'apprentissage machine s'avère concevable d'après le présent travail de recherche. Le critère qui a été utilisé dans ce travail a été un succès dans la diminution entre l'écart du *p-rule* des différents sous-groupe d'un attribut sensible. En effet, la divergence de Kullback-Leibler a été mise en oeuvre de manière à contrôler l'effet que les variables *sexe*, *âge* et l'intersection *sexe* $\times$ *âge* ont sur les prédictions des primes pures dans le Boosting Bernoulli.

Le critère d'évaluation de la discrimination choisi, le *p-rule*, a pu capter la différence entre les algorithmes sans contrainte d'équité et ceux avec la contrainte. De plus, l'ajout de poids dans la contrainte afin d'équilibrer les proportions entre les sous-groupes a permis d'obtenir de meilleurs résultats. Les primes ont ensuite été calculées en utilisant le modèle avec le λ optimal pour chaque attribut sensible. La divergence KL n'a pas été conclusive avec le modèle de Boosting Tweedie en raison de problème computationel.

Le λ optimal choisi pour chaque attribut sensible peut, cependant, être différent pour les divers sous-groupes. Il serait intéressant de mettre en oeuvre la contrainte d'équité de manière à ajouter le λ optimal propre à chaque sous-groupe.

Somme toute, ce travail vise à sensibiliser également l'industrie de l'assurance à la discrimination qui peut se trouver dans les modèles de tarification. Il est nécessaire de s'engager activement dans des actions visant à promouvoir l'égalité des chances. Toutefois, puisqu'il y a un compromis précision-équité à prendre en considération, des régulations qui s'appliquent sur tous les assureurs permettraient aussi d'éviter

les potentielles pertes financières qui peuvent découler de la mitigation de l'équité en tarification. Les régulateurs devraient développer les outils nécessaires pour assurer un traitement équitable envers les minorités visibles et, de même, définir clairement les potentielles contraintes, modèles et type d'équité que l'industrie de l'assurance pourrait éventuellement appliquer.

BIBLIOGRAPHIE

Automobile Insurance Rate Board (2022). How Auto Insurance Rates Are Calculated. Récupéré le 2023-02-28 de https://albertaairb.ca/wp-content/uploads/2022/08/How-Auto-Insurance-Rates-Are-Calculated.pdf

Barocas, S. et Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.2477899. Récupéré le 2023-02-16 de https://www.ssrn.com/abstract=2477899

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S. et Roth, A. (2017). *A Convex Framework for Fair Regression*. Rapport technique arXiv :1706.02409, arXiv. arXiv :1706.02409 [cs, stat] type : article

Bian, Y. et Zhang, K. (2023). Increasing Fairness in Compromise on Accuracy via Weighted Vote with Learning Guarantees. arXiv :2301.10813 [cs]. Récupéré le 2023-04-07 de http://arxiv.org/abs/2301.10813

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information science and statistics. New York : Springer.

Board of commissioners of public utilities (2019). Review of automobile insurance in Newfoundland and Labrador. Récupéré le 2023-02-28 de https://www.gov.nl.ca/dgsnl/files/insurance-pdf-auto-ins-review-01-29-2019.pdf

Charpentier, A. et Barry, L. (2022). L'équité de l'apprentissage machine en assurance. *10*.

Chibanda, K. F. (2022). Defining Discrimination in Insurance. p. 18.

Chouldechova, A. et Roth, A. (2018). The Frontiers of Fairness in Machine Learning. arXiv :1810.08810 [cs, stat]. Récupéré le 2023-04-04 de http://arxiv.org/abs/1810.08810

Chzhen, E., Denis, C., Hebiri, M., Oneto, L. et Pontil, M. (2020). Fair regression with Wasserstein barycenters. Dans *Advances in Neural*

Information Processing Systems, volume 33, 7321–7331. Curran Associates, Inc. Récupéré le 2023-04-04 de https://proceedings.neurips.cc/paper_files/paper/2020/hash/51cdbd2611e844ece5d80878eb770436-Abstract.html

Clark, D. R. et Thayer, C. A. (2004). Récupéré de https://www.casact.org/sites/default/files/database/dpp_dpp04_04dpp117.pdf

Co-operators (2023). L'assurance automobile par province. Récupéré le 2023-02-27 de https://www.cooperators.ca/fr-CA/resources/protect-what-matters/auto-insurance-regulation

Commission, O. H. R. *et al.* (1999). *Human Rights Issues in Insurance : Discussion Paper*. The Branch.

Dickey, M. R. (2020). Twitter and Zoom's algorithmic bias issues. Récupéré le 2023-02-28 de https://techcrunch.com/2020/09/21/twitter-and-zoom-algorithmic-bias-issues/

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. et Zemel, R. (2011). Fairness Through Awareness. *arXiv :1104.3913 [cs]*. arXiv : 1104.3913. Récupéré le 2022-05-09 de http://arxiv.org/abs/1104.3913

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C. et Venkatasubramanian, S. (2015). Certifying and removing disparate impact. arXiv :1412.3756 [cs, stat]. Récupéré le 2023-03-20 de http://arxiv.org/abs/1412.3756

Finance and Treasury Board (2004). Insurance Act. Récupéré le 2023-02-28 de https://laws.gnb.ca/en/ShowPdf/cr/2004-139.pdf

Financial Services Regulatory Authority of Ontario (2023). What determines your auto insurance rate | Financial Services Regulatory Authority of Ontario. Récupéré le 2023-02-28 de https://www.fsrao.ca/consumers/auto-insurance/understanding-auto-insurance-rates/what-determines-your-auto-insurance-rate

Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine. *Annals of statistics*, 1189–1232.

Ghosh, D. (2018). KL Divergence for Machine Learning. Récupéré le 2023-04-04 de https://dibyaghosh.com/blog/probability/kldivergence.html

Gosselin, J. (2020). La discrimination systémique, «ça existe au Québec». *La Presse*. Récupéré le 2023-02-28 de

https://www.lapresse.ca/actualites/national/2020-06-03/la-discrimination-systemique-ca-existe-au-quebec

Gouvernement du Québec (1996). Charte des droits et libertés de la personne. Récupéré le 2023-02-28 de https://www.legisquebec.gouv.qc.ca/fr/version/lc/C-12?code=se:20_1&historique=20160718

Grari, V., Charpentier, A., Lamprier, S. et Detyniecki, M. (2022). A fair pricing model via adversarial learning. *arXiv :2202.12008 [cs, stat]*. arXiv : 2202.12008. Récupéré le 2022-05-05 de http://arxiv.org/abs/2202.12008

Gursoy, F. et Kakadiaris, I. A. (2022). Error Parity Fairness : Testing for Group Fairness in Regression Tasks. arXiv :2208.08279 [cs]. Récupéré le 2023-01-20 de http://arxiv.org/abs/2208.08279

Haeri, M. A. et Zweig, K. A. (2020). The Crucial Role of Sensitive Attributes in Fair Classification. Dans *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2993–3002., Canberra, ACT, Australia. IEEE. http://dx.doi.org/10.1109/SSCI47803.2020.9308585. Récupéré le 2023-03-08 de https://ieeexplore.ieee.org/document/9308585/

Hardt, M., Price, E. et Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *arXiv :1610.02413 [cs]*. arXiv : 1610.02413. Récupéré le 2022-05-05 de http://arxiv.org/abs/1610.02413

IBM (2023). What is Boosting ? | IBM. Récupéré le 2023-05-26 de https://www.ibm.com/topics/boosting

Institut canadien des actuaires (2023). Biais et équité dans la tarification et la souscription des risques d'assurances IARD. p. 28.

Insurance Corporation of British Columbia (2012). Insurance Corporation of British Columbia.

Kingma, D. P. et Welling, M. (2022). Auto-Encoding Variational Bayes. arXiv :1312.6114 [cs, stat]. Récupéré le 2023-04-03 de http://arxiv.org/abs/1312.6114

Lichtenstein, E. (2022). Which States Ban Gender-Rating In Insurance Premiums ? | AgentSync. Section : State Regulatory Change. Récupéré le 2023-02-24 de https://agentsync.io/blog/state-regulatory-change/which-states-ban-gender-rating-in-insurance-premiums

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R. et Puri, R. (2019). Bias mitigation post-processing for individual and group

fairness. Dans *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, 2847–2851. IEEE.

Lupton, D. (2022). Machine Learning in Insurance.

Mary, J., Calauzènes, C. et Karoui, N. E. (2019). Fairness-Aware Learning for Continuous Attributes and Treatments. Dans *Proceedings of the 36th International Conference on Machine Learning*, 4382–4391. PMLR. ISSN : 2640-3498. Récupéré le 2022-06-10 de https://proceedings.mlr.press/v97/mary19a.html

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. et Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. arXiv :1908.09635 [cs]. Récupéré le 2023-04-04 de http://arxiv.org/abs/1908.09635

Milenkovic, D. (2021). Manitoba Car Insurance — All You Need to Know. Récupéré le 2023-02-27 de https://carsurance.net/canada/by-province/manitoba-car-insurance/

Mohamed, M. M. et Schuller, B. W. (2022). Normalise for Fairness : A Simple Normalisation Technique for Fairness in Regression Machine Learning Problems. arXiv :2202.00993 [cs]. Récupéré le 2023-04-04 de http://arxiv.org/abs/2202.00993

Ontario Human Rights Commission (2005). Politique et directives sur le racisme et la discrimination raciale. Récupéré le 2023-03-08 de https://www3.ohrc.on.ca/sites/default/files/attachments/Policy_and_guidelines_on_racism_and_racial_discrimination_fr.pdf

Paetzold, R. L. et Willborn, S. L. (1996). Deconstructing Disparate Impact : A View of the Model through New Lenses. *NORTH CAROLINA LAW REVIEW*, *74*.

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L. et Camps-Valls, G. (2017). Fair Kernel Learning. arXiv :1710.05578 [stat]. Récupéré le 2023-04-04 de http://arxiv.org/abs/1710.05578

Pigeon, M. (2021). Quatrième étude de cas.

Pigeon, M. (2022). Notes de cours pour MAT861A.

Reid, T. R. et report, W. P. S. W. s. r. J. S. c. t. t. (1985). Montana Implements Policy Of 'Unisex' Insurance. *Washington Post*. Récupéré le 2023-02-24 de https://www.washingtonpost.com/archive/politics/1985/10/01/montana-implements-policy-of-unisex-insurance/

c8897a22-a667-4d6e-acb2-b695702e1b31/

Seiner, J. A. (2006). Disentangling Disparate Impact and Disparate Treatment : Adapting the Canadian Approach. *Yale Law & Policy Review*, *25*(1), 95–142. Publisher : Yale Law & Policy Review, Inc. Récupéré le 2023-03-09 de https://www.jstor.org/stable/40239673

SGI (2023). Rates - SGI - SGI - Liferay DXP prd01. Récupéré le 2023-02-27 de https://sgi.sk.ca/rates

Shen, A., Han, X., Cohn, T., Baldwin, T. et Frermann, L. (2022). Optimising Equal Opportunity Fairness in Model Training. arXiv :2205.02393 [cs]. Récupéré le 2023-08-01 de http://arxiv.org/abs/2205.02393

Suresh, H. et Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Dans *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. arXiv :1901.10002 [cs, stat], http://dx.doi.org/10.1145/3465416.3483305. Récupéré le 2023-04-05 de http://arxiv.org/abs/1901.10002

Tishby, N. et Zaslavsky, N. (2015). Deep Learning and the Information Bottleneck Principle. arXiv :1503.02406 [cs]. Récupéré le 2023-04-03 de http://arxiv.org/abs/1503.02406

Tiwari, A. (2020). Insurance Risk Pricing — Tweedie Approach. Récupéré le 2023-05-12 de https://towardsdatascience.com/insurance-risk-pricing-tweedie-approach-1d71207268fc

Wang, Q., Xu, Z., Chen, Z., Wang, Y., Liu, S. et Qu, H. (2020). Visual analysis of discrimination in machine learning. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1470–1480.

Werner, G. et Guven, S. (2007). GLM Basic Modeling : Avoiding Common Pitfalls.

West, K. (2013). Gender in Automobile Insurance Underwriting : Some Insureds Are More Equal Than Others. *Alberta Law Review*, *50*(3), 679. http://dx.doi.org/10.29173/alr101. Récupéré le 2023-02-24 de https://www.albertalawreview.com/index.php/ALR/article/view/101

Yan, S., Huang, D. et Soleymani, M. (2020). Mitigating Biases in Multimodal Personality Assessment. Dans *Proceedings of the 2020 International Conference on Multimodal Interaction*, 361–369., Virtual Event Netherlands. ACM. http://dx.doi.org/10.1145/3382507.3418889. Récupéré le 2023-04-05 de https://dl.acm.org/doi/10.1145/3382507.3418889

Yang, Y., Qian, W. et Zou, H. (2016a). Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models. arXiv :1508.06378 [stat]. Récupéré le 2023-05-15 de http://arxiv.org/abs/1508.06378

Yang, Y., Qian, W. et Zou, H. (2016b). Insurance premium prediction via gradient tree-boosted tweedie compound poisson models.

Zafar, M. B., Valera, I., Rodriguez, M. G. et Gummadi, K. P. (2017). Fairness Beyond Disparate Treatment & Disparate Impact : Learning Classification without Disparate Mistreatment. Dans *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. arXiv :1610.08452 [cs, stat], http://dx.doi.org/10.1145/3038912.3052660. Récupéré le 2022-08-26 de http://arxiv.org/abs/1610.08452

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *arXiv :1511.00148 [cs, stat]*. arXiv : 1511.00148. Récupéré le 2022-05-06 de http://arxiv.org/abs/1511.00148