Anti-discrimination Insurance Pricing: Regulations, Fairness Criteria, and Models

Xi Xin, Fei Huang*

UNSW Sydney

Abstract

On the issue of insurance discrimination, a grey area in regulation has resulted from the growing use of big data analytics by insurance companies – direct discrimination is prohibited, but indirect discrimination using proxies or more complex and opaque algorithms is not clearly specified or assessed. This phenomenon has recently attracted the attention of insurance regulators all over the world. Meanwhile, various fairness criteria have been proposed and flourished in the machine learning literature with the rapid growth of artificial intelligence (AI) in the past decade, which mostly focus on classification decisions. In this paper, we introduce the fairness criteria that are potentially applicable to insurance pricing as a regression problem to the actuarial field, match them with different levels of potential and existing anti-discrimination regulations, and implement them into a series of existing and newly proposed anti-discrimination insurance pricing models, using both generalized linear models (GLMs) and Extreme Gradient Boosting (XGBoost). Our empirical analysis compares the outcome of different models via fairness-accuracy trade-off and shows their impact on adverse selection and solidarity.

Keywords: Indirect Discrimination, Fairness, AI, Big Data, Insurance Pricing

*Corresponding author: Fei Huang, feihuang@unsw.edu.au. School of Risk and Actuarial Studies, UNSW Sydney, 2052, Australia. We are grateful to Chris Dolman, Edward (Jed) Frees, and Michael Powers for valuable comments and suggestion, various seminar participants for helpful comments.

1 Introduction

In many other fields, the term 'discrimination' carries a negative connotation implying that the treatment is unfair or prejudicial, while in the insurance field, it often retains its original neutral meaning as "the act of distinguishing"[1]. Following Frees and Huang (2021), we use the word "discrimination" in an entirely neutral way, taking it to mean the act of treating different groups differently – where the groups are distinguished by salient features such as hair color, age, gender, heritage, religion, and so forth – whether such discrimination is justifiable or not.

The nature of insurance is risk pooling and the essence of pooling is discrimination, which is a business necessity for insurance companies to discriminate insureds by classifying them into different risk pools and each pool with similar likelihood of losses. Risk classification benefits insurers as it reduces adverse selection, moral hazard, and promotes economic efficiency, while high-risk consumers worry about being unfairly discriminated against by insurance companies with more frequent use of Big Data and more advanced analytics tools.

Traditionally, insurance companies are not allowed to use certain protected characteristics (those characteristics are usually also socially unacceptable) to directly discriminate policyholders in underwriting or rating, such as race, religion or national origin. Some recognized proxies for protected attributes of insureds are also restricted or even prohibited for its use in insurance practices, such as zip code, occupation or credit-based insurance score. With the rapid development of AI technologies and insurers' extensive use of Big Data, a growing concern is that insurance companies can use proxies or develop more complex and opaque algorithms to discriminate against policyholders. A grey area has resulted from this phenomenon – direct discrimination is prohibited by forbidding the use of certain factors, but indirect discrimination using proxies or more complex and opaque algorithms is not clearly specified or assessed[2]. This phenomenon has recently attracted the attention of insurance regulators all over the world[3].

Under the current anti-discrimination legal framework, some jurisdictions (e.g., EU and Australia) have defined indirect discrimination, while a similar concept of disparate impact standard is developed within the case law in the United States as a legal theory of discrimination, but the extent to which indirect discrimination or disparate impact discrimination can be restricted is still vague and undefined. In reality, a common practice is that insurance companies simply avoid using or even collecting sensitive (or discriminatory) features, and argue that the output produced by analytics algorithms without using discriminatory variables is unbiased and based only on statistical evidence (EIOPA 2019). However, indirect discrimi-

[1] See Merriam-Webster.com Dictionary, "Discriminating Among Meanings of Discrimination", available at https://www.merriam-webster.com/dictionary/discrimination.

[2] Birnbaum (2020) made a similar point in his presentation to NAIC Consumer Liaison Committee, and asked that "if discriminating intentionally on the basis of prohibited classes is prohibited – e.g., insurers are prohibited from using race, religion or national origin as underwriting, tier placement or rating factors – why would practices that have the same effect be permitted?"; Birny Birnbaum is the Executive Director of the Center for Economic Justice (CEJ).

[3] We believe this is a legal term derived from U.S. employment discrimination laws and is synonymous with intentional discrimination.

nation may still occur when proxy variables (i.e. identifiable proxy) or opaque algorithms (i.e. unidentifiable proxy) are used. Therefore, there is an urgent need globally for insurance regulators to propose standards to identify and address the issues of indirect discrimination including algorithmic discrimination.

Machine learning experts are devoted to the discussion of algorithmic bias and fairness by introducing various fairness criteria, and most of these criteria broadly fall into two main categories: individual fairness criteria and group fairness criteria. Intuitively by their names, these fairness criteria aim to either achieve fairness at the individual or group level and an inevitable conflict exists between group fairness and individual fairness, see also Binns (2020). In general, most of previous fairness literature focuses on a classification problem or decision and its application in employment, education, lending and criminal justice, etc. However, there is little research on insurance applications, particularly on insurance pricing as a regression problem, see Lindholm et al. (2022) and Vincent et al. (2022).

Although insurance discrimination draws more and more attention in recent years, for example see Frees and Huang (2021) and Dolman and Semenovich (2019), there is little research on the relationship between different insurance regulations, fairness criteria, and pricing models. Understanding their inter-relationship, however, is important both for practicing actuaries (to implement appropriate models in practice) and governments (to understand the impact of different regulations and design auditing tools). To this end, this paper aims to establish the linkage among insurance regulations, fairness criteria, and insurance pricing models. In particular, this paper reviews anti-discrimination laws and regulations of different jurisdictions with a special focus on indirect discrimination of the general insurance industry. We introduce the fairness criteria that are potentially applicable to insurance pricing as a regression problem to the actuarial field, match them with different levels of potential and existing anti-discrimination regulations, and implement them into a series of existing and newly proposed anti-discrimination insurance pricing models, using both generalized linear models (GLMs) and Extreme Gradient Boosting (XGBoost). Our empirical analysis compares the outcome of different models via fairness-accuracy trade-off and shows the impact on customer behavior and solidarity. In particular, we demonstrate the appealing potential of anti-discrimination pricing models for rate making compared to common industry practice (fairness through unawareness).

The rest of the paper is organised as follows. In Section 2, we examine and compare anti-discrimination laws and regulations in the insurance industry with a focus on general insurance (auto insurance and home insurance), by reviewing several major insurance markets such as the United States, the European Union and Australia. We also summarize the current efforts to deal with algorithmic discrimination and various reasons for supporting or opposing insurance discrimination. In Section 3, we summarize different fairness criteria originating from machine learning area and establish a connection with the legal and regulatory frameworks examined in Section 2. In Sections 4, we summarize existing and newly proposed anti-discrimination insurance pricing models and match them with fairness criteria in Section 3. In Section 5, we evaluate and compare different anti-discrimination insurance pricing methods to remove (or reduce) indirect discrimination based on a real general insurance dataset from the perspectives of both group fairness and individual fairness. Section 6 summarizes different regulations to

mitigate indirect discrimination and match them with individual or group fairness criteria and representative models that directly satisfy the regulations. Section 7 concludes the paper.

2 Laws and Regulations on Insurance Discrimination

In this section, we will examine and compare anti-discrimination laws and regulations in the insurance industry with a focus on general insurance (auto insurance and home insurance) by reviewing existing laws and regulations in several major jurisdictions. Because not all jurisdictions have anti-discrimination regulations on insurance discrimination, we mainly review regulations in the United States, the European Union and Australia. We also summarize the trends of current efforts on future laws and regulations to deal with algorithmic discrimination in the era of big data, and various reasons about why insurance companies discriminate in practice.

2.1 Prohibited Features and Direct Discrimination

- **Direct Discrimination.** Direct discrimination occurs when a person is treated less favourably than another person simply because one of their protected characteristics is not the same. If the person's corresponding risk factor is not used by insurers, such discrimination can be completely avoided.

Direct discrimination refers to the direct use of a protected attribute that is determined by law and prohibited from being used as a risk factor, also known as disparate treatment[4]. Common protected attributes include race, national or ethnic origin, religion or belief, gender, sexual orientation, age and disability, which usually vary by jurisdiction, line of business and even different insurance stages.

In the United States, insurance anti-discrimination laws and regulations vary greatly by state and a comprehensive comparison is provided in Avraham, Logue, and Schwarcz (2014b) for fifty-one jurisdictions by focusing on five lines of insurance and each comparing nine different characteristics as of 2012. Commonly, the issue of insurance discrimination may be covered in a broader anti-discrimination legal framework. In the European Union, Directive 2004/113/EC (a.k.a. "Gender Directive") and Directive 2000/43/EC (a.k.a. "Racial Equality Directive") prohibit direct (and indirect) insurance discrimination on the grounds of gender and racial or ethnic origin. Both Directives as EU law only sets Union-wide minimum level of standard for the protection against discrimination and most member states offer broader protection under national law (European Commission 2014). In Australia, federal anti-discrimination laws cover a wide range of grounds broadly including race, sex, disability and age, and insurers are given exemptions and allowed to discriminate in certain circumstances (Australian Law Reform Commission 2003).

[4]We believe this is a legal term derived from U.S. employment discrimination laws and is synonymous with intentional discrimination.

2.2 Indirect Discrimination

- **Indirect Discrimination.** After avoiding direct discrimination[5], indirect discrimination occurs when a person is still treated unfairly than another person by virtue of implicit inference from their protected characteristics, based on an apparently neutral practice such as using proxy variables from the non-protected characteristics of policyholders (i.e. identifiable proxy), or opaque algorithms (i.e. unidentifiable proxy).

Regulators and other stakeholders often reach a common understanding of indirect discrimination. Indirect discrimination is expressly defined in the anti-discrimination laws of various jurisdictions (e.g. the European Union and Australia), which usually include the following essential elements: 1) caused by a facially neutral practice, policy or rule that applies to everyone in the same way; 2) related to a protected characteristic specified in law; 3) individuals with a certain protected characteristic are treated unfairly or disproportionately compared with those who do not share it. A parallel definition – disparate impact discrimination originated in the United States and was initially proposed in the field of employment. Its definition is considered to include all the basic elements of indirect discrimination. We believe that disparate impact is a subset of indirect discrimination and only intends to cover unintentional discrimination[6].

However, in the insurance field, the current regulation on indirect discrimination is mainly through prohibiting or restricting the use of certain proxies for protected features. Some traditionally or recently recognized proxy variables, such as zip code, credit information, education level, and occupation, are regulated mainly because of their negative impact on (racial) minorities and low-income individuals. In the United States, insurers are prohibited or severely restricted to use drivers' education and occupation in automobile insurance rating in at least four states (Consumer Reports 2021). To the extent of our knowledge, there is no existing legal framework in any jurisdiction to explicitly assess indirect discrimination in the insurance sector. Miller (2009) commented, "thus far no court has actually applied the disparate impact (or adverse impact) standard to insurance rates, but it is only a matter of time before some court does so". We refer interested readers to Appendix A for a detailed discussion on the evolvement of U.S. insurance discrimination regulations, including the disparate impact standard and its applicability in the insurance industry.

2.3 Algorithmic Discrimination and Responses to Big Data

Algorithmic discrimination refers to the biased outcomes or decisions produced by algorithms and is usually considered as a subset of indirect discrimination. In the Big Data Analytics (BDA) thematic review conducted by the European Insurance and Occupational Pensions Authority (EIOPA 2019) based on 222 participated motor or health insurers from 28 European

[5]Note that we use a narrow definition of indirect discrimination assuming that the law has prohibited or will prohibit direct discrimination on protected characteristics, and we limit the scope of our research on indirect discrimination to this situation. We recognize that direct discrimination and indirect discrimination on the same protected characteristic may occur simultaneously, but if direct discrimination is allowed, then the provisions on indirect discrimination will be meaningless.

[6]Although it may cover intentional indirect discrimination that is too difficult to prove discriminatory intent under a disparate treatment case.

jurisdictions, 31% of insurance firms already actively used BDA tools and another 24% of firms plan to use them within the next three years, and these new data analytics tools are generally used on pricing and underwriting, claims management and sales and distribution, whereas insurers have only taken limited approaches to ensure fair and ethical outcomes in the use of BDA in underwriting and pricing[7]. Xenidis and Senden (2019) explore algorithmic discrimination in the era of big data within the current EU legal framework.

Insurance regulators are publicly seeking advice on algorithmic discrimination issues. In the United States, the National Association of Insurance Commissioners (NAIC) published guiding principles on artificial intelligence (AI)[8] in August 2020 including a key principle "encouraging industry participants to take proactive steps to avoid proxy discrimination against protected classes when using AI platforms[9]" developed by the NAIC's Big Data and Artificial Intelligence Working Group. However, the term "proxy discrimination" has not yet been defined by the NAIC (see Prince and Schwarcz (2019) for exploring the definition of proxy discrimination in the age of big data) and it is unclear for insurers on how to comply with the guiding principles to avoid proxy discrimination in practice. In the European Union, EIOPA established a Consultative Expert Group on Digital Ethics in Insurance as a follow up of the thematic review and assists to develop digital responsibility principles in insurance regarding fairness and ethical issues that arise with the use of digital technologies in practice. In Australia, the Australian Human Rights Commission (AHRC) published a technical paper on addressing the issue of algorithmic bias when using Artificial intelligence (AI) in decision making (AHRC 2020).

2.4 Why Do Insurance Companies Discriminate?

There is no simple answer to this question and different factors are taken into consideration. Frees and Huang (2021) focus on discrimination in the insurance context and assess the appropriateness of insurance discrimination by reviewing social and economic principles. Avraham, Logue, and Schwarcz (2014a) explain variations in insurance anti-discrimination laws in the U.S. among states, characteristics and lines of coverage by considering three efficiency or fairness properties that U.S. state legislatures seek to balance: predictive capacity, adverse selection and illicit discrimination (see also Wortham (1986b) and Gaulding (1994)). Loi and Christen (2021) provide an ethical analysis of insurance discrimination in private insurance by relating philosophical moral arguments to the discussion of fair predictive

[7]EIOPA (2019) notes that "some insurance firms declared that they 'smoothed' the output of such algorithms, for instance by not using machine learning without human intervention or by establishing caps to the outputs of these tools in order to ensure ethical outcomes (e.g. not charging vulnerable customers excessively).... Regarding the potential difficulties to access insurance for high-risk consumers,... motor insurance firms also referred to already existing mechanisms in some jurisdictions such as insurability schemes or the obligation of insurance firms to not reject motor third-party liability insurance (MTPL) consumers (albeit there is no limit in maximum premium)...."

[8]See National Association of Insurance Commissioners (NAIC) Principles on Artificial Intelligence (AI), available at https://content.naic.org/sites/default/files/inline-files/AI%20principles%20as%20Adopted%20by%20the%20TF_0807.pdf, as a response to the OECD Principles on Artificial Intelligence.

[9]See NAIC Unanimously Adopts Artificial Intelligence Guiding Principles, available at https://content.naic.org/article/news_release_naic_unanimously_adopts_artificial_intelligence_guiding_principles.htm#:~:text=Washington%20(August%2020%2C%202020),safe%2C%20secure%20and%20robust%20outputs.

algorithms in the machine learning field.

In terms of insurance practices, insurance companies can usually be exempted from using certain protected factors, or if they can show that the use of these factors is actuarially justified. Insurance premiums should reflect the the expected losses of the insured risk based on the principle of actuarial fairness. For more details about actuarial fairness, Landes (2015) reviews how the principal of actuarial fairness is formulated within the insurance industry. Meyers and Van Hoyweghen (2018) analyse how actuarial fairness has been enacted in different ways in insurance practice over time, from the traditional fair discrimination to contemporary behavioural-based fairness, the latter is based on the support of personalized data, such as personal driving style or lifestyle.

An opposite and somewhat ambiguous concept is solidarity, which is commonly related to social insurance, and the principle of solidarity emphasizes the sharing of risks across groups, even if the use of the risk-rating factor can be actuarially justified, see Lehtonen and Liukko (2011) for summarizing different forms of insurance solidarity. A well-known example is the unisex rule in the European Union, which prohibits gender-specific premium differentiation and covers private insurance contracts: in the *Test-Achats* ruling[10], the European Court of Justice (ECJ) ruled Article 5(2) of Directive 2004/113/EC was invalid – the controversial clause "permits proportionate differences in individuals' premiums and benefits where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data" – and consequently, insurers are no longer able to use gender as a risk factor (i.e., no exemption is permitted) to determine premiums or benefits of insurance services from December 21, 2012 and individual insurance policies should be issued at gender neutral rates.

In general, there is an inevitable conflict between insurance companies and high-risk consumers on the degree of strictness of insurance discrimination regulation. Insurers statistically discriminate between policyholders according to individual risk profile in order to treat similar policyholders similarly – focusing on personalization or individualization of insurance products based on the principle of actuarial fairness. On the contrary, high-risk consumers, with the support of consumer advocates and some regulators, welcome strict regulation (e.g., promote the application of disparate impact standard in the U.S insurance industry) to better protect their interests and avoid discrimination – focusing on standardization of insurance products based on the principle of solidarity. This also reflects the different views of insurance in different jurisdictions and lines of business – whether it is regarded as economic commodity or social good (Frees and Huang 2021), see Section 6 for different regulation examples.

2.5 Existing Regulations and Discussions about Future Regulations

Insurance discrimination definitions can be wrapped into different names. Chibanda (2022) summarizes various terms that are used in defining discrimination by different stakeholders in the U.S. insurance industry (including unfair discrimination, proxy discrimination, disparate

[10]European Court of Justice (ECJ) (2011) Judgement of the Court, Case C-236/09, available at https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A62009CJ0236.

treatment and disparate impact) and finds that these terms either focus on inputs or effects. In fact, most existing regulations focus on inputs by prohibiting or restricing the use of certain attributes. The most popular effects-oriented regulatory example is the unisex regulation in the EU as described in Section 2.4 – EU insurers are compulsory to provide the same premium or benefit for men and women given the same profile of individuals, while gender is still allowed to be used as long as it does not result in individual differences in premiums or benefits. Another recent example is the Colorado Senate Bill 21-169 in the United States, which was passed and signed into law in July 2021, and its definition of unfair discrimination has a "disparate impact" component, which could be the first insurance regulation to focus on the effects of discrimination at the group level – that is common in other areas such as lending, housing or college admissions. Please refer to Appendix A for detailed explanation of this Colorado Senate Bill.

3 Fairness Criteria for Insurance Pricing

Extensive research has been conducted in the field of machine learning to combat discrimination in big data and artificial intelligence. For various reasons, most researchers tend to define the notion of fairness and propose measures to achieve fairness accordingly, rather than define the notion of discrimination (or unfairness) and develop methods to prevent or mitigate discrimination. In this section, we will examine and discuss fairness criteria that are applicable in the context of insurance pricing.

Most of the existing fair machine learning literature is related to employment or housing discrimination due to the disparate impact provisions (i.e. see Section 2.2) contained in several U.S. federal laws and hence focuses on binary classification decisions, such as hiring or lending. Barocas and Selbst (2016) analyse the instances of discriminatory data mining under Title VII jurisprudence for employment discrimination taking into account both disparate treatment and disparate impact theories of liability and provide a bridge between computer science literature and existing anti-discrimination laws and regulations in employment decisions. Hutchinson and Mitchell (2019) study fairness and unfairness definitions from 1960s in the fields of education and employment and connect to machine learning fairness criteria. Binns (2018) link fair machine learning with extant literature in moral and political philosophy. Berk et al. (2018) integrate existing research in criminology, computer science and statistics to address both fairness and accuracy for risk assessments in criminal justice settings. However, there has been little research linking fairness criteria proposed in the machine learning literature to actuarial pricing applications and this section will fill this gap.

We provide a list of notions below, which will be used for fairness definitions in the insurance pricing context.

- Let an ordered triple $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space, where $\mathbb{P}$ represents the real world measure.
- Let X_P denote the protected attribute; for simplicity, we let X_P be a categorical variable and has only two groups $X_P = \{a, b\}$, $X_P = a$ is the advantaged group and $X_P = b$ is the disadvantaged group.

- Let X_{NP} denote other available (non-protected) attributes, and hence the feature space is $X = \{X_P, X_{NP}\}$.
- Let $\hat{Y}$ denote the predictor or the decision outcome of interest, $\hat{Y} \in \mathbb{R}$. In our context, $\hat{Y}$ is the premium charged by the insurer, and in this paper, we assume that $\hat{Y}$ is approximately equal to the pure premium and ignore any expenses or profit loadings.
- Let Y denote the observed outcome of interest, $Y \in \mathbb{R}$. Note that Y is not known when the policy is issued, Y is a measure of real claim experience observed by the insurer over a given period after policy issuance, and in theory, Y corresponds to the revised premium that the insurance company would charge if the insurer had known the insured's the actual claim settlement experience over a given period in advance.

3.1 Individual Fairness and Group Fairness

As early as the 1970s, research in other fields has noticed the conflict between individual fairness and group fairness, see Thorndike (1971) and Sawyer, Cole, and Cole (1976). In particular, Sawyer, Cole, and Cole (1976) distinguish individual parity and group parity as follows:

> a conflict arises because the success maximization procedures based on individual parity do not produce equal opportunity (equal selection for equal success) based on group parity and the opportunity procedures do not produce success maximization (equal treatment for equal prediction) based on individual parity.

In the insurance field, individual fairness is analogous to the idea of treating similar people similarly (see Dwork et al. (2012), Kusner et al. (2017) and Zemel et al. (2013)), while group fairness aims to ensure group level fairness across all groups in X_P by treating individuals (differently) with regard to the (protected) group they belong to – here we adopt the broader meaning of group fairness, although the term is sometimes used interchangeably with demographic parity (Definition 5 in Section 3.3) in the field of machine learning.

This classical trade-off is reflected in the views of insurance companies and high-risk consumers (or regulators) on insurance discrimination regulations. Insurers support risk-based pricing based on statistical discrimination which is close to the principle of individual fairness to treat similar people similarly. Conversely, consumer representatives for high-risk individuals (i.e., consumer advocates, regulators) seek to protect the interests of low-income or racial minority individuals, who support the use of group-level fairness criteria to avoid disparate impact against the protected class. This also reflects the different views of insurance, economic commodity or social good, in different jurisdictions and lines of business (Frees and Huang 2021).

In terms of insurance regulations, the current insurance regulation pays more attention to individual fairness rather than group fairness; and in practice, prohibiting the use of a protected characteristic as the most common anti-discrimination regulatory method corresponds to the fairness notion of fairness through unawareness. Moreover, the actuarial principle that defines unfairly discriminatory insurance rates is similar to the concept of individual fairness – treating similar risks similarly and not treating similar risks differently. Based on a different

motivation, the movement to introduce disparate impact standard into the insurance industry aims to achieve parity across groups based on a protected feature (e.g., race or gender) in order to protect minority groups in insurance practices, and an extreme case in practice is community rating in health insurance, which ensures group fairness on all features and everyone pays the same premium. See Section 6 for the summary of regulations and the matching between different regulations and fairness criteria.

In the following subsections, we will introduce fairness criteria by individual fairness and group fairness, respectively. Although it is generally difficult to impose both individual and group fairness criteria in a method at the same time, targeting to meet an individual fairness criterion does not mean group fairness criteria cannot be moderately satisfied under certain conditions, and vice versa. These individual and group fairness criteria to be introduced in Sections 3.2 and 3.3 are usually not mutually exclusive.

3.2 Individual Fairness Criteria

> **Definition 1 – Fairness Through Unawareness:** fairness is achieved if the protected attribute X_P is not explicitly used in calculating the insurance premium $\hat{Y}$.

Satisfying Definition 1 is a sufficient condition to avoid direct discrimination on the basis of the protected attribute X_P by prohibiting the use of X_P in rating, and the same premium will be offered across different groups of X_P if non-protected attributes X_{NP} are the same. Definition 1 assumes that the premiums will be fair if insurers are unaware of protected attributes in rate making, whereas this assumption is generally unrealistic because protected attributes are often correlated with other non-protected attributes in the insurance data and indirect discrimination may still persist via other attributes that are proxies of the protected attribute, and therefore produce unfair outcomes to protected groups.

Definition 1 is commonly used as a baseline approach due to its apparent simplicity in machine learning and it is also the default scenario for insurers in practice because they are often not allowed to collect certain sensitive variables. For example, EU insurers usually choose not to collect sensitive protected variables such as race, ethnic origin and gender (EIOPA 2019), and similarly, US insurers generally do not know the race, religion, or national origin of the insureds (NAMIC 2020).

> **Definition 2 – Fairness Through Awareness:** a predictor $\hat{Y}$ satisfies fairness through awareness if it gives similar predictions to similar individuals (Dwork et al. 2012; Kusner et al. 2017).

Definition 2 is originally proposed by Dwork et al. (2012) as a concept of individual fairness in classification and aims to overcome the unfairness to individuals under group fairness criteria, and its notion is based on the idea that similar people should be treated similarly. Importantly, a task-specific distance metric is required to measure the similarity between the individuals considering human insight and domain information (Dwork et al. 2012) and similar individuals should receive a similar distribution over outcomes, and hence the difficulty or the limitation in applying this definition is to find a proper similarity metric

in a given context (Kim, Reingold, and Rothblum 2018). In subsequent research based on the idea of Definition 2, Zemel et al. (2013) introduce a fair classification algorithm aiming to achieve both group fairness and individual fairness (i.e., statistical parity and fairness through awareness) and Berk et al. (2017) encodes fairness as a family of flexible regularizers spanning from group fairness to individual fairness covering intermediate or hybrid fairness notions for regression problems.

Hardt (2013) points out that insurance risk metrics are practical examples of their work on fairness through awareness. For example, insurance scores or credit-based insurance scores are used to help insurers in underwriting or pricing, typically in automobile and homeowners insurance. These numerical ratings are based on consumers' credit information and indicate how an individual manages its financial affairs, because they are often good indicators of insurance claims (III 2019).

> **Definition 3 – Counterfactual Fairness:** a predictor $\hat{Y}$ is counterfactually fair for an individual if "its prediction in the real world is the same as that in the counterfactual world where the individual had belonged to a different demographic group". (Kusner et al. 2017; Wu, Zhang, and Wu 2019), or mathematically, given that $X = x$ and $X_P = a$, for all y and for simplicity X_P has only two groups $\{a, b\}$, a predictor $\hat{Y}$ is counterfactually fair if
>
> $$\mathbb{P}(\hat{Y}_{X_P \leftarrow b}(U) = y \mid X_{NP} = x, X_P = b) = \mathbb{P}(\hat{Y}_{X_P \leftarrow a}(U) = y \mid X_{NP} = x, X_P = b)$$

Following Kusner et al. (2017), let U denote relevant unobserved latent or exogenous variables (e.g., driving habits data can be potentially collected by insurance telematics), and $\hat{Y}_{X_P \leftarrow b}$ is interpreted as the value of $\hat{Y}$ if X_P had taken value b (Pearl and others 2000). The notion of counterfactual fairness is introduced in Kusner et al. (2017) based on causal methods and it is an individual-level definition, and in their paper, Kusner et al. (2017) also contrast their fairness criteria with individual fairness or group fairness (i.e., Definitions 2 and 5). Counterfactual fairness is referred to as counterfactual demographic parity in Barocas, Hardt, and Narayanan (2019) due to its close similarity to Definition 6.

A similar work based on causal reasoning was proposed independently by Kilbertus et al. (2017) at about the same time, and two causal discrimination criteria are defined after introducing the concepts of resolving variables and proxy variables. In the subsequent development, Chiappa (2019) introduces a novel notion of path-specific counterfactual fairness for complicated scenarios by only correcting the causal effect of the protected attribute on the decision along the unfair pathways (not fair pathways). Di Stefano, Hickey, and Vasileiou (2020) indicate the lack of research on incorporating causality into popular discriminative machine learning models. For more details about causality and discrimination, we refer readers to Chapter 4 of Barocas, Hardt, and Narayanan (2019). Despite the popularity of counterfactual fairness as a promising technique since it was proposed, Kasirzadeh and Smart (2021) argue that "even though counterfactuals play an essential part in some causal inferences, their use for questions of algorithmic fairness and social explanations can create more problems than they resolve."

The advantage of these causal fairness criteria is to focus on the role of causality in fairness reasoning. To interpret Definition 3 in the insurance pricing scenario, a predictive model is used to decide the premium $\hat{Y}$, the premium charged of an individual from the disadvantaged group $X_P = b$ remains the same if this person had been from the advantaged group $X_P = a$, we can ascertain that this person has been treated fairly under the concept of counterfactual fairness. Kusner et al. (2017) provide three ways of achieving counterfactual fairness, and the simplest way to make $\hat{Y}$ counterfactually fair is to use only the observable non-descendants of X_P.

> **Definition 4 – Controlling for the Protected Variable:** As definied in Defition 6 by Lindholm et al. (2022), a *discrimination-free price* for Y w.r.t X_{NP} is definied by
>
> $$h^*(X_{NP}) := \int_{x_P} \mathbb{E}[Y|X_{NP}, x_P] d\mathbb{P}^*(x_P),$$
>
> where $\mathbb{P}^*(x_P)$ is definied on the same range as $\mathbb{P}(x_P)$.

Driven by concerns over the proxy effects of X_{NP} on X_P, Definition 4 (or the procedure based on Definition 4) was proposed that aims to decouple the protected attribute X_P from non-protected attributes X_{NP}, see Pope and Sydnor (2011) and Lindholm et al. (2022). The discrimination-free price based on Definition 4 is acquired by averaging best-estimate prices $\mathbb{E}[Y|X_{NP}, x_P]$ (or `Model 1`'s prediction outputs as labelled in Section 4) over protected attributes using $\mathbb{P}^*(d)$, and a simple choice $\mathbb{P}^*(x_P) = \mathbb{P}(x_P)$ is recommended in Lindholm et al. (2022), which is justified by them using causal inference arguments.

3.3 Group Fairness Criteria

Barocas, Hardt, and Narayanan (2019) classified most of the group fairness criteria in the classification setting into three categories: independence ($\hat{Y} \perp X_P$), separation ($\hat{Y} \perp X_P \,|\, Y$) and sufficiency ($Y \perp X_P \,|\, \hat{Y}$), and Barocas, Hardt, and Narayanan (2019) comment that these fairness criteria are all observational because they are properties of the joint distribution of $\{X_{NP}, X_P, \hat{Y}, Y\}$ compared with the non-observational fairness criteria discussed earlier (e.g., causal fairness criteria). Although observational fairness criteria have inherent limitations (Kilbertus et al. 2017; Barocas, Hardt, and Narayanan 2019) such as indistinguishability, these criteria are appealing because of their ease of use.

As mentioned earlier in this section, due to the ambiguity in defining the actual outcome Y in the insurance domain, we will focus on fairness criteria in the independence category, or in other words, demographic parity and its variants in this subsection.

> **Definition 5 – Demographic Parity (or Statistical Parity):** a predictor $\hat{Y}$ satisfies demographic parity if
>
> $$\mathbb{P}(\hat{Y}|X_P = a) = \mathbb{P}(\hat{Y}|X_P = b)$$

Demographic parity, also known as statistical parity or group fairness, is the most basic fairness criterion to achieve group fairness (i.e., the broader meaning of group fairness, as

defined in Section 3.1). The criterion requires that the predictor $\hat{Y}$ and the protected attribute X_P are statistically independent, and ensure fairness to be achieved at the group level across groups a and b. For regression, a similar definition of statistical parity is defined based on the cumulative distribution function in Agarwal, Dudik, and Wu (2019).

In the insurance environment, satisfying demographic parity implies the average premium will be approximately the same across groups a and b ($\mathbb{E}(\hat{Y}|X_P = a) = \mathbb{E}(\hat{Y}|X_P = b)$), and cross-subsidy usually exists between insureds under demographic parity. Since the disadvantaged demographic group ($X_P = b$) generally corresponds to the group of high-risk insureds to the insurance company, this criterion implies that low-risk insureds will cross-subsidize high-risk insureds and inevitably, the insureds will be treated differently based on their protected attribute X_P, and therefore a disadvantage of this criterion is that we treat all groups similarly without considering the potential differences across groups (Caton and Haas 2020).

> **Definition 6 – Disparate Impact (the Four-Fifths Rule):** a predictor $\hat{Y}$ has no disparate impact if the following ratio is above than a certain threshold τ (Feldman et al. 2015):
>
> $$\frac{\mathbb{P}(\hat{Y} = \hat{y}|X_P = b)}{\mathbb{P}(\hat{Y} = \hat{y}|X_P = a)} > \tau$$

There are approximate versions of demographic parity (Barocas, Hardt, and Narayanan 2019), and Definition 6 can be seen as a more flexible approximate version of demographic parity. The expression of this definition focuses on the concerns of severe disparate impact on the disadvantaged group ($X_P = b$), which represents a socially protected group or a minority group that is often (unfairly) discriminated against. As a relaxation of demographic parity criterion, we accept the deviation of the two conditional probabilities within a predetermined threshold. In the United States, the well-known "80 percent" rule (or the four-fifths rule) regarding employment discrimination in the hiring process is obtained if τ is set to 0.8, and $\hat{Y}$ is the positive outcome – the applicant is accepted. The "80 percent" rule was codified in the 1978 *Uniform Guidelines for Employee Selection Procedures*[11], advocated by the U.S. Equal Employment Opportunity Commission (EEOC), which is intended to detect adverse impact (i.e., disparate impact) on a protected group in employee selection procedures. Currently, the four-fifths rule is often used along with more sophisticated statistical methods.

In insurance rate making, since a higher $\hat{Y}$ indicates a worse outcome for policyholders and presumably premiums of the disadvantaged groups ($X_P = b$) are higher than the advantaged group ($X_P = a$), we need to adjust the above inequality as follows:

$$\frac{\mathbb{P}(\hat{Y} = \hat{y}|X_P = a)}{\mathbb{P}(\hat{Y} = \hat{y}|X_P = b)} > \tau$$

When $\tau = 0.8$, we will get the corresponding four-fifths rule on insurance pricing. Compared

[11] See available at https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml; and questions and answers from EEOC website https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines

with Definition 5, variations from demographic parity are allowed in Definition 6, which takes into account the potential differences between groups in X_P and sets allowable premium differentiation through τ to limit the influence of severe disparate impact against the disadvantaged class $X_P = b$. In practice, this definition may implicitly assumes that insurance companies are allowed to use the protected attribute X_P but the impact of X_P is restricted within a predetermined range. For example, under the Affordable Care Act (ACA), age rating ratio shall not exceed 3:1 using a 21-year-old as the baseline and tobacco rating ratio for tobacco users shall not exceed 1.5:1, and each state can request a rating ratio lower than the federal standard.

> **Definition 7 – Conditional Demographic Parity (or Conditional Statistical Parity):** a predictor $\hat{Y}$ satisfies conditional demographic parity if
>
> $$\mathbb{P}(\hat{Y}|X_{NP_{legit}} = x_{NP_{legit}}, X_P = a) = \mathbb{P}(\hat{Y}|X_{NP_{legit}} = x_{NP_{legit}}, X_P = b)$$
>
> where $X_{NP_{legit}}$ denotes a subset of "legitimate" attributes within unprotected attributes in the feature space ($X_{NP_{legit}} \subseteq X_{NP} \subset X$) that are permitted to affect the outcome of interest (Corbett-Davies et al. 2017; Verma and Rubin 2018).

Conditional demographic parity (or conditional statistical parity) is also a relaxation of demographic parity criterion. Fairness is achieved at the group level across groups a and b after controlling for a set of "legitimate" attributes L. Moreover, like Definition 1, this definition does not strictly reduce disparities across groups in X_P after permitting a set of legitimate attributes, and Corbett-Davies et al. (2017) state that conditional demographic parity "mitigates these limitations of the blindness approach while preserving its intuitive appeal", and therefore as a better alternative of Definition 1. Similarly, the idea of legitimate variables is used in Kilbertus et al. (2017) by introducing the concepts of non-resolving and resolving variables in casual reasoning.

Under conditional demographic parity, while aiming to maintain group fairness to avoid disparate impact against minority individuals, insurance regulators are more flexible to either approve some rating factors that are allowed to cause disparities among groups in X_P, or restrict other rating factors that may act as proxies of X_P. In general, the criterion of conditional demographic parity provides more flexibility to insurance companies as a copromise between fairness through awareness and demographic parity.

Note that Definition 1 is a special case of Definition 7 if all attributes are legitimate and Definition 5 is a special case of Definition 7 if all attributes are non-legitimate. Definition 7 is also similar to the *Actuarial Group Fairness* definition proposed by Dolman and Semenovich (2019). They are equivalent when the legitimate variables are the variables used for cost modeling and Y denotes the market premium. The conditional demographic parity definition is also consistent with the unfair discrimination definition provided in the recent Colorado Senate Bill 21-169 when the legitimate variables are the variables used for cost/losses modelling, see Appendix A for more discussions on this Bill.

The expression of Definition 7 can be extended to a more flexible version similar to that of Definition 6 to Definition 5, which we call conditional disparate impact, that is written as

follows:

$$\frac{\mathbb{P}(\hat{Y} = \hat{y} | X_{NP_{legit}} = x_{NP_{legit}}, X_P = b)}{\mathbb{P}(\hat{Y} = \hat{y} | X_{NP_{legit}} = x_{NP_{legit}}, X_P = a)} > \tau$$

Insurance regulators allow group-level premium differences caused by legitimate predictors, and limit those caused by non-legitimate predictors within a predetermined range.

3.4 Further Works and Practical Limitations of Insurance Fairness Criteria

For future research, an interesting question to consider is whether insurance fairness criteria would benefit from considering actual outcome or observed outcome of interest. A major improvement of demographic parity is equalized odds proposed by Hardt, Price, and Srebro (2016) (or separation criterion in Barocas, Hardt, and Narayanan (2019)) requiring $\hat{Y} \perp X_P | Y$, and for a binary classification decision, this criterion is equivalent to ensuring the same true positive rates and false positive rates across the demographic groups a and b. The use of Y is critical in equalized odds, which is the outcome observed at a later point in time after the corresponding decision $\hat{Y}$ is made. However, Y may not reflect the "true type" particularly where Y contains a significant element of chance, as in the case for insurance claims, see Dolman and Semenovich (2019).

4 Anti-discrimination Insurance Pricing Models

In this section, we propose several anti-discrimination pricing strategies to eliminate or reduce indirect discrimination based on the insurance fairness criteria discussed in Section 3, and we also explore how these strategies correspond to existing or potential anti-discrimination statutes as discussed in Section 2. For each anti-discrimination pricing strategy, we can further categorize them into pre-processing (on the training data prior to modelling), in-processing (during model training) and post-processing (on the outputs after modelling) methods based on the implementation time of each fairness criterion at different modelling stages.

For this study, we show each model in its simplest form as a linear model for illustration purpose and use the same notation as in Section 3: the rating variables X can be split into protected variables (X_P) and non-protected variables (X_{NP}) and Y represents our response variable, which can also be interpreted as claim counts or claim amounts in addition to pure premiums in Section 3, and $\hat{Y}$ represents the predicted value of Y. An empirical analysis using both GLM and XGBoost is conducted later in Section 5 and all model labels in this section (`Model 1`, `Model 2`, `Model 3`, `Model 4` and `Model 5`) correspond to the same model labels in Section 5. The models we considered in this sections are linked to the different fairness criteria defined in Section 3. In particular, `Model 2`, `Model 3`, `Model 4`, and `Model 5` correspond to `Definition 1`, `Definition 5`, `Definition 7`, and `Definition 4`, respectively.

4.1 Model 1: Full Model

In the full model, all attributes can be used and `Model 1` allows both direct and indirect discrimination on grounds of all protected characteristics in our dataset, which can be expressed as

$$\hat{Y}_{M1} = f_1(X_{NP}, X_P).$$

Here f_1 is some fixed but unknown function of X. The baseline model or the full model (`Model 1`)'s linear representation is $\hat{Y}_{M1} = \mathbf{1}\ b_{0,1} + X_P\ b_{1,1} + X_{NP}\ b_{2,1}$.

4.2 Model 2: Excluding Protected Variables

Extending from `Model 1`, `Model 2` is fit using only non-protected variables to avoid direct discrimination that can be expressed as

$$\hat{Y}_{M2} = f_2(X_{NP}).$$

`Model 2`'s linear representation is $\hat{Y}_{M2} = \mathbf{1}\ b_{0,2} + X_{NP}\ b_{2,2}$.

`Model 2` corresponds to the notion of *Fairness Through Unawareness*, see the discussion of `Definition 1` in Section 3. `Model 2` undoubtedly avoids indirect discrimination unless $X_{NP} \perp X_P$.

4.3 Model 3: Fitting with Debiased Variables

In this paper, we apply pre-processing methods to achieve *Demographic Parity* (`Definition 3`) by fitting with unbiased data that aims to remove the dependence between X_{NP} and X_P, because $X_{NP} \perp X_P$ is a sufficient condition for $\hat{Y} \perp X_P$. Let $X^\star_{NP}$ denote the debiased version of non-protected predictors after removing their dependence with X_P. `Model 3` can be expressed as

$$\hat{Y}_{M3} = f_3(X^\star_{NP}).$$

Its linear representation is $\hat{Y}_{M3} = \mathbf{1}\ b_{0,3} + X^\star_{NP}\ b_{2,3}$.

4.3.1 Method 1: Using Disparate Impact (DI) Remover

For the first method, we use the Disparate Impact (DI) Remover as detailed in Feldman et al. (2015). Given a protected variable X_P and a single continuous or ordinal non-protected variable X_{NP}, the conditional distribution of X_{NP} given $X_P = x_P$ is defined as $X_{NPx_P} = Pr(X_{NP}|X_P = x_P)$. The cumulative distribution function of X_{NPx_P} is denoted as F_{X_P} and the corresponding quantile function is denoted as $F^{-1}_{X_P}$.

Define A as a "median" distribution of its quantile function, which is expressed as follows:

$$F^{-1}_A : F^{-1}_A(u) = \text{median}_{x_P \in X_P} F^{-1}_{X_P}(u)$$

and therefore, the adjusted non-protected predictor $X^\star_{NP}$ is found by

$$x^{\star}_{NP} = F_A^{-1}(F_{X_P}(x_{NP}))$$

where the resulting $X^{\star}_{NP}$ is fair and strongly preserves rank within groups.

4.3.2 Method 2: Using Orthogonal Predictors

For the second method, we use orthogonal predictors by pre-adjusting each non-protected attribute in X_{NP} to be uncorrelated with the protected attribute (X_P) as first proposed in Frees and Huang (2021) for insurance applications. We regress each of the non-protected variables in X_{NP} onto all protected variables X_P:

$$X_{NP} = \mathbf{1}\ b_{0,OP} + X_P b_{1,OP}$$

and let $\bar{X}_{NP}$ denote the predicted value of X_{NP}, $X^{\star}_{NP}$ is found as $X^{\star}_{NP} = X_{NP} - \bar{X}_{NP}$.

The second method only removes all linear correlation between X_P and X_{NP} and does not guarantee that X_{NP} after transformation is mutually independent of X_P. Therefore, this method satisfies demographic parity criterion (Definition 5 in Section 3) when assuming there is only linear dependence between X_P and X_{NP}. Interestingly, if the protected attribute X_P as a subset of X, is also the parent (or direct cause) of random variables X_j in X in a causal model and strong level 3 assumption in Kusner et al. (2017) is met, `Model 3` further satisfies counterfactual fairness criterion (Definition 3 in Section 3).

4.3.3 Model 3 in Practice

In general, direct discrimination is avoided like `Model 2` and indirect discrimination is reduced or removed by making each non-protected attribute neutral on the protected attribute. `Model 3` will ensure the average premium charged is approximately the same across demographic groups by satisfying Definition 5. In insurance applications, `Model 3` can avoid disparate impact on members of a protected class that may constitute discrimination within the U.S. legal framework, and guarantee that insurers will not be subject to disparate impact liability, as discussed later in Section 6.1. However members of the previously advantaged group may find themselves disadvantaged, which coincides with the classic trade-off between group fairness and individual fairness, as discussed in Section 3.1.

The limitations of this approach include the inevitable lose of information when adjusting X and the failure to remove (potentially discriminatory) interactions effects in X_P when considering multiple protected attributes (Berk 2009; Berk et al. 2018). A more complicated alternative that seeks to minimize information loss in X is proposed in Johndrow and Lum (2019). Also, `Model 3` requires knowledge of protected variables in both the training phase and the prediction phase, and in particular, the essence of `Model 3` is to use information about the protected characteristics of policyholders to transform the non-protected variables to be discrimination-free in insurance rating, however, insurers may face obstacles in practice to access to information of protected variables and potential concerns from policyholders.

4.4 Model 4 - Fitting with Legitimate and Debiased Non-legitimate Variables

We propose a new model, labelled `Model 4`, which is a compromise between `Model 2` and `Model 3` and will satisfy *Conditional Demographic Parity* (`Definition 7`). X_{NP} is further split into legitimate variables $X_{NP_{legit}}$ and non-legitimate variables $X_{NP_{not}}$. `Model 4` allows disparities in insurance premium across protected groups through pre-determined legitimate variables ($X_{NP_{legit}}$) in X_{NP}, while other attributes in X_{NP} are transferred ($X^{\star}_{NP_{not}}$) using bias mitigation methods as described in `Model 3`. Because $X_{NP_{not}} \perp X_P$ is a sufficient condition for $\hat{Y} \perp X_P \mid X_{NP_{legit}}$, conditional demographic parity criterion (Definition 7 in Section 3) is achieved under `Model 4`, which is expressed as

$$\hat{Y}_{M4} = f_4(X^{\star}_{NP_{not}}, X_{NP_{legit}}).$$

Its linear representation is $\hat{Y}_{M4} = \mathbf{1}\ b_{0,4} + X^{\star}_{NP_{not}}\ b_{2,4} + X_{NP_{legit}}\ b_{3,4}$.

`Model 4` is proposed as a more flexible alternative to `Model 3` and this approach also achieves group fairness but allowing flexibility through legitimate attributes compared with `Model 3`. In the insurance field, insurance regulators can determine that certain attributes are legitimate and then allow group-level premium differences between protected demographic groups to come from these legitimate variables, and therefore `Model 4` can play a more important role in practice.

4.5 Model 5 - Controlling for the Protected Variable

`Model 5` is consistent with the methods provided in Pope and Sydnor (2011) and Lindholm et al. (2022), and will satisfy `Definition 4`. As a post-processing approach, this method was originally proposed in Pope and Sydnor (2011), where it was formally presented and thoroughly evaluated in a linear regression setting, while this approach can be seamlessly integrated into models with more complex structures. This model begins by estimating the full model (`Model 1`) to obtain the coefficient estimates and then averages across the values of the protected variable in the population for predictions. `Model 5` is expressed as

$$\hat{Y}_{M5} = \frac{1}{N}\sum_{j=1}^{N} \hat{f}_1(X_{NP}, X_P = x_{pj}),$$

where N denotes the number of policyholders, x_{pj} denotes the value (vector) of the protected variables for the j^{th} policyholder, and $\hat{f}$ denotes the estimated `Model 1`. `Model 5`'s linear representation is $\hat{Y}_{M5} = \mathbf{1}\ b_{0,1} + \bar{X}_P\ b_{1,1} + X_{NP}\ b_{2,1}$, where the coefficients $b_{0,1}$, $b_{1,1}$, and $b_{2,1}$ are from the full model (`Model 1`) and $\bar{X}_P$ is the average value (vector) of the protected variables for the population. Protected attributes X_P are only used in the training phase, while in the prediction phase, we average out X_P using population average statistics or sample average estimates ($\bar{X}_P$) in determining individual pure premiums. Pope and Sydnor (2011) believe this approach will allocate the appropriate relative weight to each fitted predictor reflecting

its true predictive power by sacrificing part of the model's accuracy[12], Lindholm et al. (2022) extend Pope and Sydnor's research and provide a rigorous probabilistic justification of this discrimination-free pricing procedure, and additionally they propose several ways to mitigate potential pricing bias at portfolio level.

`Model 5` also achieves fairness at the individual level as a better alternative to `Model 2`, and it tends to address the issue in `Model 2` when the protected characteristics of policyholders X_P are omitted, proxy variables in X_{NP} will have increased predictive power driven by their ability to proxy for X_P, which is restricted by fitting both X_P and X_{NP} in the model in order to restrict the inference of X_{NP} from X_P.

5 Empirical Analysis

5.1 French Dataset and Its Background

In this section, we analyze a dataset from a French private motor insurance drawn from the `R` package `CASdatasets` (Dutang, Charpentier, and Dutang (2015)) – `pg15training`, which was used for the first pricing game organized by the French institute of Actuaries in 2015. The training dataset (`pg15training`) contains 100,000 third-party liability (TPL or civil liability) policies[13] observed from 2009 to 2010 including the guarantee for two types of compensation – material damage (e.g., damage to a building or another vehicle) and bodily injury that could be caused to a third party when the driver is held responsible for an accident, and this simple guarantee is the mandatory minimum guarantee as required by law[14]. In the following analysis, we narrow the scope of our work to focus on third-party material claims, where such claims were filed more frequently than third-party bodily injury claims in our dataset.

We adopt the frequency-severity approach and two methods are used for a comparison of different method types – one is the standard frequency-severity GLM approach using Poisson regression and gamma regression, and the other is built on Extreme Gradient Boosting (XGBoost) models using Poisson deviance loss for claims frequency and gamma deviance loss for claims severity. XGBoost was proposed by Chen and Guestrin (2016) as a novel gradient tree boosting method and has rapidly gained popularity due to its high efficiency in computational speed and predictive performance in applications in many fields. In terms of insurance claim prediction, XGBoost outperforms other methods at handling large training data and many missing values (Fauzan and Murfi 2018). For all XGBoost models, we perform a grid search for tuning hyperparameters by steps using five-fold cross-validation, and we refer interested readers to Fauzan and Murfi (2018) for detailed grid search scheme. For details of the GLM and XGBoost models, please refer to Appendix B.

[12]It also highlights the classical trade-off between model accuracy and model fairness in fair modelling discussion. The authors (Pope and Sydnor 2011) further believe this approach will potentially produce more economically efficient outcomes for society.

[13]In `pg15training`, the first 21 records have been removed because they are duplicate records, which have non-zero claim count (`Numtppd`) and zero claim amount (`Indtppd`). After removal, there are exactly 50000 policies each year in 2009 and 2010.

[14]See the Directorate of Legal and Administrative Information, available at https://www.service-public.fr/particuliers/vosdroits/F2628;

We consider those anti-discrimination pricing models introduced in Section 4 to address indirect gender discrimination, using the same labels from `Models 1` to `5`. `Gender` is the protected variable in our empirical analysis[15], and we use the following non-protected explanatory variables X_{NP}: `Age`[16], `Bonus`, `Group1` (car group), `Density` (the density of inhabitants), `Value` (car value), `Insurance Score`[17]. Our response variable is pure premium (frequency$\times$severity), and each individual's predicted pure premium is adjusted to correct for portfolio level bias for GLM `Models 3` to `5` and all XGBoost models on the basis of GLM `Model 2` by proportionally adjusting each individual's premium according to its pre-adjusted predicted value (Lindholm et al. 2022).

Moreover, following Frees and Huang (2021) we develop an artificial gender proxy[18] for the probability of being female for each driver, which takes into account ten moderately efficient gender proxy variables[19] that are simulated independently using the gender information of each observation, and this gender proxy is created based on the idea that the accumulation of some medium proxy variables will form a strong gender proxy. Although it may constitute indirect discrimination in the EU under the Gender Directive, or intentional discrimination in the United States, this artificial proxy predictor is added to `Model 2`, `Model 3` and `Model 4`, leading to `Model 2'`, `Model 3'` and `Model 4'` respectively.

5.2 Disparate Impact Remover and How It Works on Age?

The Disparate Impact (DI) remover is applied on all predictors in `Model 3` and all non-legitimate predictors in `Model 4`. Among all predictors, we note its effect on age stands out compared to other predictors. By sub-grouping individuals by gender and age, we find that younger people are at greater risk than older people, and men are at greater risk than women at each age, while excluding `Gender` in modelling, women in aggregate are at greater risk than men because the proportion of women is relatively higher at younger ages.

DI remover aims to remove the effect of gender on age, and in general, men's ages are adjusted downward and women's ages are adjusted upward as displayed in Figure 2. One important property of the DI remover is that it strongly preserves rank within male or female policyholders. However, there is a question of legitimacy here – whether adjusting the age of policyholders is a reasonable action. Alternative methods to remove disparate impact include

[15]Effective from December 21, 2012, EU insurance companies are completely prohibited to charge different premiums on the basis of gender after the *Test-Achats* ruling disallowed actuarial exemptions on the use of gender as described in Section 2.4, and similarly, seven states in the United States currently forbid the use of gender as a rating factor in auto insurance.

[16]Instead of binning into classes, `Age` is fit in a continuous function form in GLMs using the approach in Schelldorfer and Wuthrich (2019).

[17]We create an insurance score for each policyholder using `Type` (car type), `Category` (car category), `Occupation`, `Group2` (region of the driver home) and `Age`.

[18]Alternatively, gender proxies can be constructed based on variables in the training sample only (`Age` is highly influential in the gender proxy in this example). However, in this case the developed proxy is ineffective for `Model 2'`, as there is no new information added to `Model 2'` compared to `Model 2`.

[19]We simulate five male binary proxy variables and five female binary proxy variables. For example, in order to simulate the male proxy variable, given the gender of a person, each male has a 60% chance of being in the positive class, while each female only has a 40% chance.

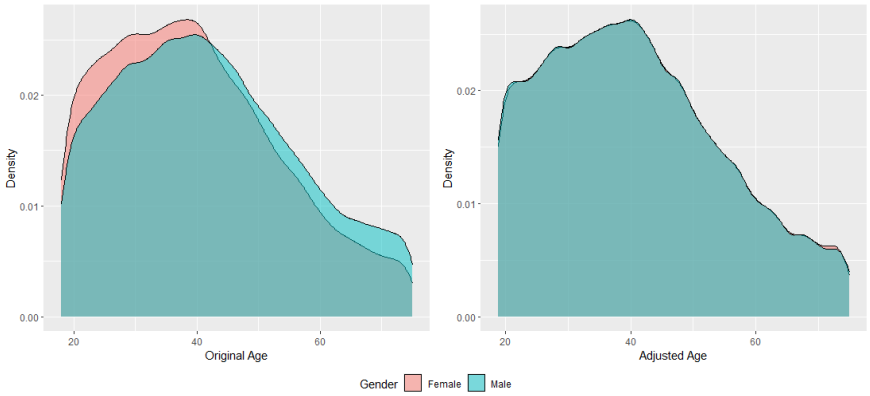

Figure 1: Probability Density Plots of Age by Gender Before and After Adjusting for Age Using the DI Remover

reweighting and resampling (Kamiran and Calders 2012).

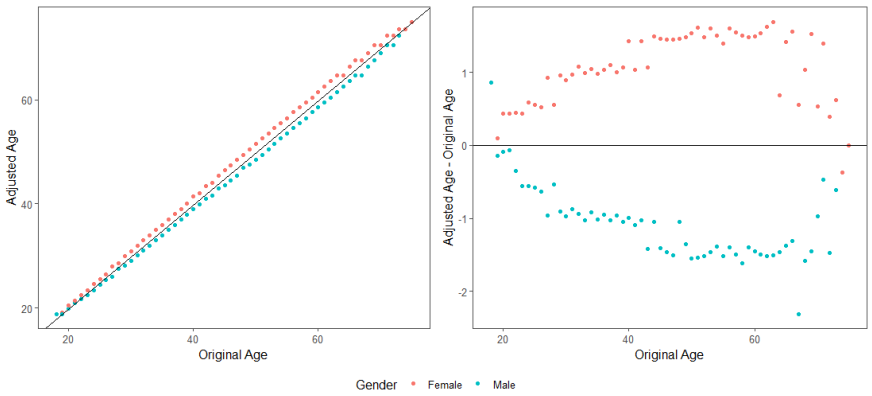

Figure 2: The Effect of DI Remover by Age and Gender

5.3 Model Comparison

Table 1 displays the means of fitted pure premiums by gender for each pair of model and method, which helps us understand the model performance in terms of group fairness, see Section 3.3. In general, the GLM and XGBoost methods provide similar results in means for each model. `Model 1` as the baseline model displays the biggest discrepancy between male and female distributions. After excluding the use of gender, `Model 2` is the only fitting procedure that does not require to collect gender in both training and prediction phases, and interestingly, women pay higher premiums than men on average as they are on average younger than men in the dataset. `Model 3` achieves the demographic parity criterion and group fairness is ensured, while the difference in means almost disappears after removing the influence of gender in all other predictors as we expected. `Model 4` is a promising insurance anti-discrimination model as a compromise between `Models 2` and `3`, and by introducing legitimate variables, we allow for deviation of group fairness from these predictors. As a better alternative to `Model 2` that also focuses on individual fairness, `Model 5` performs

	Model 1	Model 2	Model 2'	Model 3	Model 3'	Model 4	Model 4'	Model 5
GLM Male	130.47	114.03	118.53	117.38	120.78	115.26	119.40	113.95
GLM Female	95.66	124.05	116.25	118.23	112.34	121.90	114.73	124.18
XGBoost Male	131.06	114.41	118.45	117.74	120.63	116.59	119.97	114.24
XGB Female	94.50	123.39	116.37	117.61	112.60	119.61	113.74	123.68

Table 1: Comparison of Means of Predicted Pure Premiums by Model, Method and Gender After Portfolio Level Adjustment

similarly to `Model 2` when there is no strong gender proxy in the training data. In general, `Model 2`, `Model 2'` and `Model 5` meet the EU unisex premium standard, that is, the same auto insurance premium will be charged to male and female drivers given the same driver profile.

To compare the performance of different methods and models, we use Root Mean Square Error (RMSE) and normalized Gini index as our model evaluation metrics. Root Mean Square Error is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

For a sequence of numbers $\{s_i, ..., s_n\}$, we denote $r(s_i) \in \{1, ..., n\}$ as the rank of s_i in the sequence in an increasing order, and the normalized Gini index is defined as (Ye et al. 2018)

$$\text{Normalized Gini Index} = \frac{\frac{\sum_{i=1}^{n} y_i\, r(\bar{y}_i)}{\sum_{i=1}^{n} y_i} - \sum_{i=1}^{n} \frac{n-i+1}{n}}{\frac{\sum_{i=1}^{n} y_i\, r(y_i)}{\sum_{i=1}^{n} y_i} - \sum_{i=1}^{n} \frac{n-i+1}{n}}$$

and therefore the normalized Gini index utilizes pure premium predictions ($\bar{y}_i$) only through their relative orders, and a larger normalized Gini index indicates better model predictions.

In parallel, we also introduce a model fairness measure to indicate how well each model performs on group fairness inspired by Definition 6 and we call this *disparate impact ratio*, which is defined as follows:

$$\text{Disparate Impact Ratio} = \frac{\mathbb{E}(\hat{Y} = \hat{y}|X_P = b)}{\mathbb{E}(\hat{Y} = \hat{y}|X_P = a)}$$

To approximate the insurance version of the four-fifths rule, we expect this fairness score to be in the range of 0.8 to 1.2. In other words, we hope the difference in premiums on average between groups a and b do not deviate too much.

5.3.1 Scenario Analysis: Group Fairness and Prediction Accuracy

In total, we consider four different scenarios to show the effects of anti-discrimination pricing models with respect to group fairness (demographic parity) and prediction accuracy. Three scenarios (Scenarios 1-3) are created by choosing different legitimate predictors in `Model 4`,

and additionally an extra scenario (Scenario 4) is created by adding an additional gender proxy to all models to make `Model 5` more differentiable compared to `Model 2`.

- **Scenario 1**: let `Insurance Score` be the only non-legitimate predictor in `Model 4`, we consider Scenario 1 as our baseline scenario;
- **Scenario 2**: let both `Insurance Score` and `Density` be non-legitimate in `Model 4`;
- **Scenario 3**: let `Age` be the only non-legitimate variable in `Model 4`;
- **Scenario 4**: an artificially created `Gender Proxy` is added in all models and let the `Gender Proxy` be the only non-legitimate predictor in `Model 4`;

The Fairness-Accuracy plots are shown in Figures 3-6. In `Models 3` and `4`, we pre-adjust all or some of the non-protected predictors using the DI remover to make them gender-neutral by removing their dependence with `Gender` and we note that adjusting an individual predictor may either improve or reduce the accuracy and fairness of the model, and overall, the effect of adjusting for `Age` or `Insurance Score` is positive due to improved fairness and accuracy, while the effect is negative for `Density` due to decreased accuracy[20].

In Scenarios 1 to 3, `Models 2` and `5` perform similarly and their model performance are different only when there is at least one moderate gender proxy in the training data, so we add an artificially created gender proxy in all models in Scenario 4. It can be noticed that the effect of this gender proxy is different when using the GLM and XGBoost methods. Our empirical analysis shows that the XGBoost method is generally more sensitive to small gender-related differences, and we suggest insurance regulators or practitioners need to be aware that different pricing methods may have different degrees of sensitivity to the protected variable. This finding echoes the recommendation given in the EIOPA (2019) report that EU regulators consider the option of introducing specific governance requirements for specific BDA tools and algorithms.

In Scenario 4, `Models 4` and `5` perform similarly in terms of fairness, while `Model 4` outperforms `Model 5` (especially for XGBoost) slightly in terms of accuracy when the gender proxy is introduced in the data. We also notice that `Models 4` and `5` perform similarly in terms of both accuracy and fairness when there is a strong gender proxy in the training set.

Overall, Model 1 has the best prediction accuracy and worst group fairness in all scenarios. XGBoost has better prediction accuracy compared to GLM. Since Models 2-5 all satisfy the four-fifth rule according to Definition 6 in all scenarios, we could select the best model based on their prediction accuracy. In particular, XGBoost Model 4 achieves the best trade-off in scenarios 1 and 3. XGBoost Models 2 and 5 (Model 3 comes the second) achieve the best trade-off in scenario 2. XGBoost Model 2 (Model 4 comes the second) achieves the best trade-off in scenario 4.

[20] `Insurance Score`: In Scenario 1, we compare `Model 2` with `Model 4`; `Density`: we compare Model 4's performance between Scenarios 1 and 2; `Age`: In Scenario 3, we compare `Model 2` with `Model 4`;

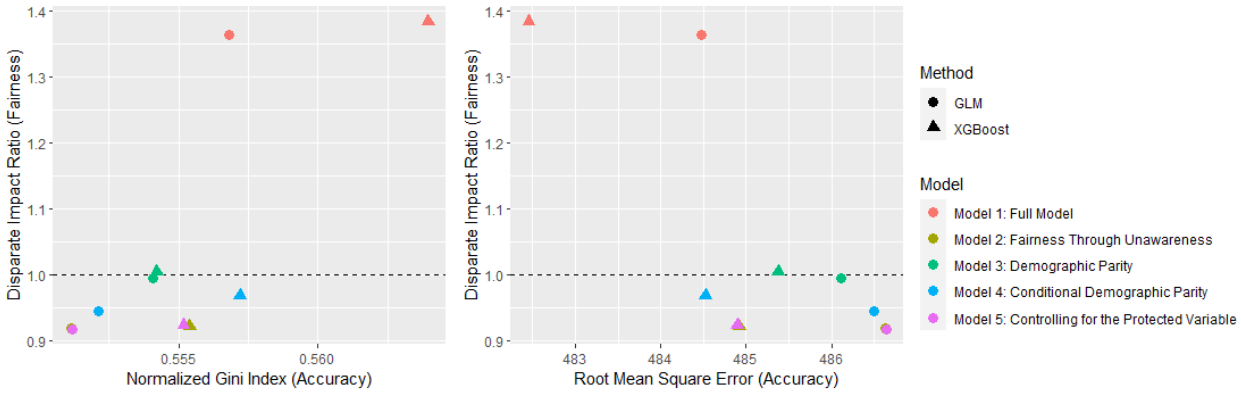

Figure 3: Fairness-Accuracy Plot (Scenario 1)

Figure 4: Fairness-Accuracy Plot (Scenario 2)

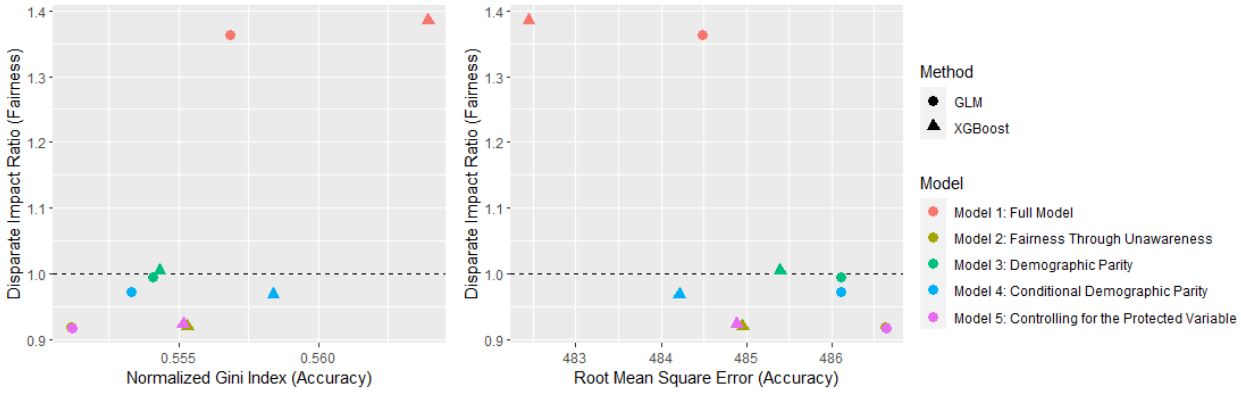

Figure 5: Fairness-Accuracy Plot (Scenario 3)

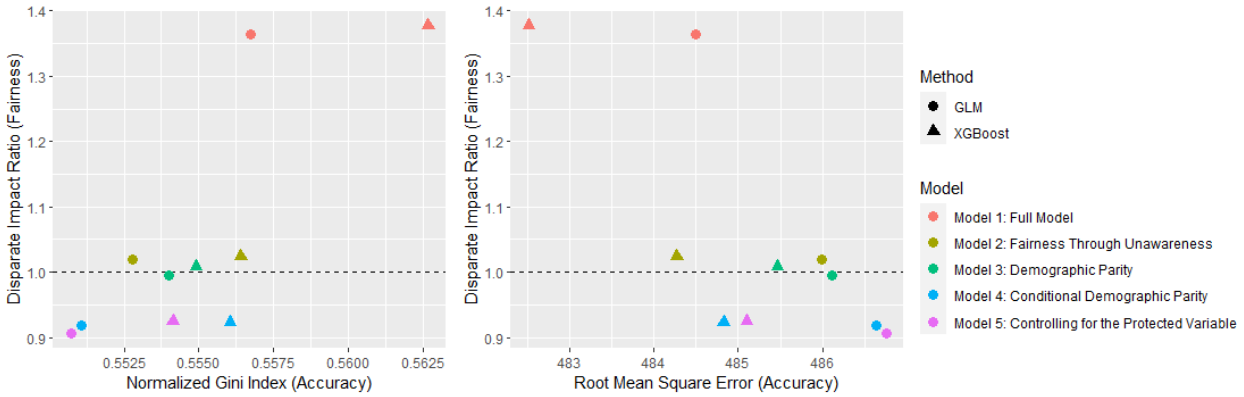

Figure 6: Fairness-Accuracy Plot (Scenario 4)

5.3.2 Adverse Selection and Solidarity

Insurance practitioners may be concerned that the use of fair models will harm the principle of actuarial (individual) fairness and lead to adverse selection as a result of implementing anti-discrimination pricing models. Following Goldburd et al. (2016), double lift charts are drawn to compare the relative performance of two models. We analyse adverse selection and consumer behavior by comparing fair models with the common unawareness practice, and the competitor (benchmark) model is assumed to be GLM `Model 2`. We use Scenario 1 as an example for illustration purposes throughout this section. The steps of creating a double lift chart is as follows:

1. We find the pure premium ratio for each individual based on a pair of benchmark and fair models, and sort the ratios from lowest to highest. Pure premium ratio is defined as a comparison of predicted pure premiums of one model to another model for the same policyholder.

$$\text{Pure premium ratio} = \frac{\text{Predicted Premium of Benchmark Model}}{\text{Predicted Premium of Fair Model}}$$

2. We create bins of equal volume exposure based on the pure premium ratios calculated.

3. For each bin, we calculate the average predicted premium for each model and the average actual experience based on actual claims.

Double lift charts are created separately by gender to compare the rating plans of the benchmark model to a fair model. The first and last bins from each double lift chart represent the two models that disagree with each other the most, and when an insurance company switches from an unawareness model to a fair model, it is most likely to lose customers from the first bin and gain customers from the last bin. In general, the effects of adverse selection using fair models compared to using the benchmark model are limited and the benefit of fairness may occur. For example, comparing GLM `Model 3` with GLM `Model 2` in Figure 7, insurers implementing a fair model will lose relatively high-risk male customers and gain low-risk male customers. And this pattern reverses for females. In Figure 8, we see that the

overall pattern is closer to the male one, that is insurers implementing a fair model will lose relatively high-risk customers and gain low-risk customers. The difference between GLM `Model 4` and GLM `Model 2` is relatively small, see Figures 11 and 12 in Appendix C.

In addition, if the fair model is based on XGBoost (see Figures 13-16 in Appendix C), we notice that insurers implementing XGBoost `Model 3` and XGBoost `Model 4` will always lose relatively high-risk male customers and gain low-risk customers for both genders and the premium difference from the benchmark GLM `Model 2` is larger using XGBoost fair models than using GLM fair models. This is consistent with our findings earlier that XGBoost models have better forecasting performance than GLM models.

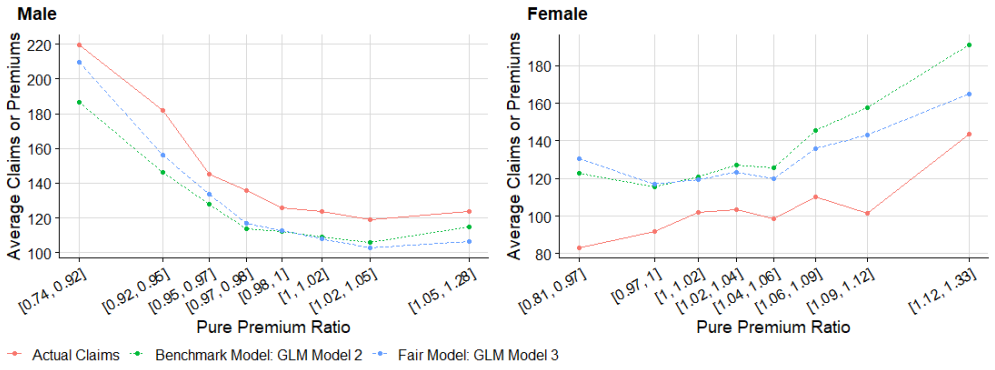

Figure 7: Double Lift Charts By Gender (GLM Model 3 vs. GLM Model 2)

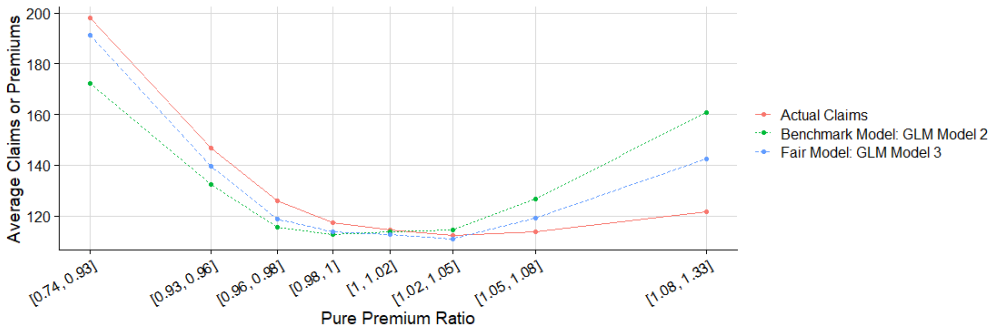

Figure 8: Double Lift Chart (GLM Model 3 vs. GLM Model 2)

Considering adverse selection is from the perspective of the insurance company, in contrast, the concept of solidarity views insurance products from the perspective of the shared responsibility in a community. Insurance products may seek an effective balance between customer segmentation and risk pooling (Henckaerts et al. 2021). Following Henckaerts et al. (2021), we assess the principle of solidarity by comparing the relative and average premium difference of each model by age and gender with respect to the benchmark models, including GLM `Model 2`, GLM `Model 1`, and actual claims costs, see Figures 9-10 and Figures 17-20 in Appendix C. Let $\bar{Y}_{fair,i}$ and $\bar{Y}_{bench,i}$ denote the average predicted premium of the fair model and benchmark model for age i, respectively. We define Relative Premium Difference and Average Premium Difference for age i below:

$$\text{Relative Premium Difference for age } i = \frac{\bar{Y}_{fair,i} - \bar{Y}_{bench,i}}{\bar{Y}_{bench,i}},$$

$$\text{Average Premium Difference for age } i = \bar{Y}_{fair,i} - \bar{Y}_{bench,i}.$$

`Model 3`, which focuses on group fairness, is theoretically the best model for the solidarity principle. It is clear from Figure 9 that, compared to GLM `Model 2`, the main subsidy of GLM `Model 3` is from males to females (except for older ages) in order to ensure group fairness among genders, and mostly between young people in terms of dollar amount (see the rhs of Figure 9). However, compared to GLM `Model 1` (see Figure 10), all other GLM fair models have the subsidy from females to males. Similar patterns can be observed when using XGBoost as fair models, see Figures 17 and 18 in the Appendix C. When using the actual claims cost as the benchmark model, we find that the patterns are more volatile across ages (see Figures 19 and 20 in Appendix C) fluctuating around zero for most ages, and young females subsidy young males for most models. It is also interesting to see that XGBoost Model 1 has the smallest average premium difference compared to the actual claims cost over all ages, showing that it is the model provides the most accurate risk estimates.

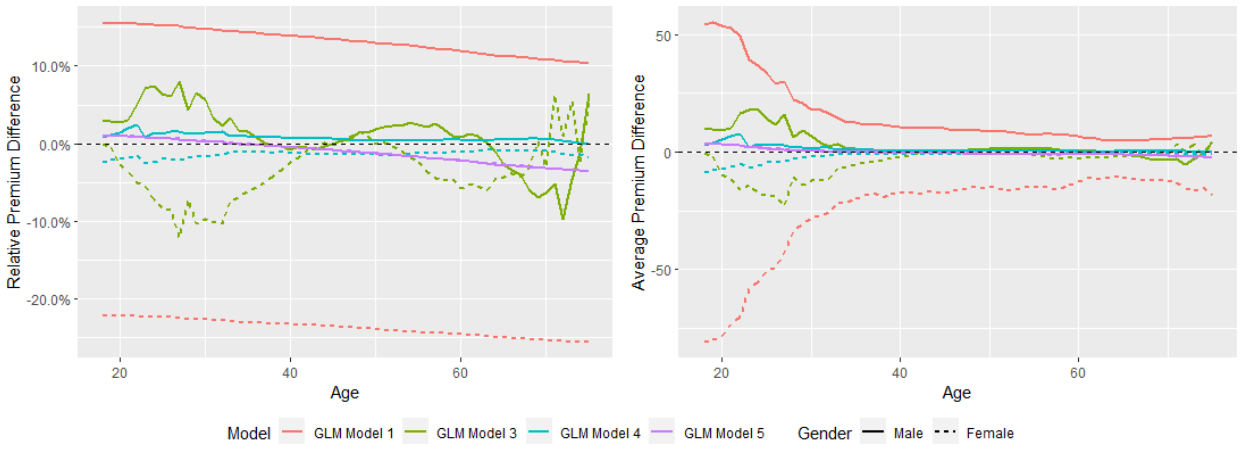

Figure 9: Relative and Average Premium Difference (GLM Models vs. GLM Model 2)

6 Regulation Comparison

In this section, we will compare different regulations on indirect discrimination and the order of the various existing or potential regulations on insurance discrimination is roughly based on the strictness of regulations, from the least restrictive "no regulation" to the most restrictive "community rating". Although we will discuss from the perspective of insurers or insurance regulations, the practical examples of regulations that will be discussed are not limited to the field of insurance. As discussed in Frees and Huang (2021), the extent of insurance rate regulation varies by jurisdiction and by line of business, which reflects different views of insurance - whether it is regarded as economic commodity or social good.

Figure 10: Relative and Average Premium Difference (GLM Models vs. GLM Model 1)

6.1 Different Regulations

6.1.1 No Regulation

At one extreme, insurance companies are free to adjust premiums using any factors, without any restrictions or prohibitions, and without prior approval from regulatory agencies, and this situation can also refer specifically to a certain variable.

6.1.2 Restriction on the Use of a Protected Variable

Insurers can be restricted on the use of a protected variable. If this variable is an important rating factor and allowed to be used, regulators can limit its impact on compressing total premium ranges between the high-risk group and the low-risk group. For example in the United States, under the Affordable Care Act (ACA), age rating ratio shall not exceed 3:1 using a 21-year-old as the baseline and tobacco rating ratio for tobacco users shall not exceed 1.5:1. Each state can request a rating ratio lower than the federal standard.

6.1.3 Prohibition on the Use of a Protected Variable

By prohibiting the use of a certain variable, direct discrimination on that characteristic is not allowed by laws or regulations, and starting from this regulation, we shift our focus to the mitigation or elimination of indirect discrimination on a protected characteristic. In addition, the direct consequence of such prohibition is that individuals from different protected groups should be offered the same premiums and benefits on the same insurance policy given the same profile on other rating factors, and the prohibited protected variable is generally still allowed to be used in rating at the aggregate level (e.g., `Model 5`) if the individual level data of such variable is available to insurers.

As a well-known example of anti-discrimination legislation, insurance companies in the EU are not allowed to use gender as a risk-rating factor in insurance products and should offer mandatory unisex premiums and benefits at the individual level. Long before, in the United States, the state of Montana has implemented unisex insurance legislation on insurance

premiums and benefits for all types of insurance since 1985, but it is also the only state while several other states have failed to introduce similar anti-discrimination legislation (that is, for all types of insurance).

6.1.4 Restriction on the Use of a Proxy Variable

Assume that direct discrimination on a protected variable is prohibited, insurers can be further restricted by regulators on the use of a certain proxy variable as a surrogate of the protected feature, and such restrictions can help prevent unfair or discriminatory practices by insurance companies to attract low-risk groups based on a protected characteristic of individuals by lowering their premiums (or raising premiums to exclude high-risk groups).

For example, if all insurers use the same rating regions allocated by the regulator, the impact of indirect discrimination caused by redlining can be partially resolved. In Australia, New South Wales is divided into five geographical zones or rating regions designated by State Insurance Regulatory Authority and insurers providing NSW Compulsory Third Party (CTP) insurance are not allowed to differentiate further (e.g. via postcode) on the basis of locality within a designated geographical zone. Also, under the ACA, each state needs to divide up the areas of the state by establishing uniform geographic rating areas based on counties, three-digit zip codes, or metropolitan statistical areas for all health insurance issuers in the individual and small group markets[21] (but insurers are not compulsory to operate in all areas in a state, see Fang and Ko (2018)).

6.1.5 Prohibition on the Use of a Proxy Variable

Insurers can be prohibited directly from using certain proxy variables to protect their negative impact on racial minorities or low-income individuals, such as zip code, credit information, occupation, education level, employment status and so on.

6.1.6 Disparate Impact Standard

In the United States, disparate impact claims are cognizable under three federal statutes concerning employment or housing discrimination: Title VII of the Civil Rights Act of 1964 (Title VI), the Age Discrimination in Employment Act of 1967 (ADEA) and the Fair Housing Act of 1968 (FHA); after three landmark U.S. Supreme Court rulings on disparate impact for each Act. In particular, on June 25, 2015, the U.S. Supreme Court held that disparate impact claims are cognizable under the Fair Housing Act in the landmark decision of *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.* (2015)[22], and it is believed[23] that disparate impact rule can be applied to prove unfair

[21] https://www.cms.gov/CCIIO/Programs-and-Initiatives/Health-Insurance-Market-Reforms/state-gra

[22] Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. 519 (2015).

[23] Before that, on February 8, 2013, the U.S. Department of Housing and Urban Development (HUD) issued a final rule entitled "Implementation of the Fair Housing Act's Discriminatory Effects Standard" ("the 2013 rule") that authorizes disparate impact claims under the Fair Housing Act (FHA) as a formal interpretation of the Act, consistent with HUD's long-held view. In particular, HUD restated its position that the Fair Housing Act applies to homeowners insurance, and hence disparate impact standard is applicable to prohibit

discrimination allegations with respect to home insurance.

In the college admission context, a classic question is that whether it is fair to use standardized test scores (e.g., SAT or ACT) to measure students from different ethnic groups. Some U.S. colleges and universities as recipients of federal funds should avoid disparate impact discrimination based on race or sex in the admissions or scholarship process under Title VI of the Civil Rights Act of 1964 (i.e., discrimination on the basis of race, color, or national origin) and Title IX of the Education Amendments of 1972 (i.e., discrimination on the basis of sex), see the report produced by the University of California (2008). Similarly, in the employment context, race norming as "the practice of converting individual test scores to percentile or standard scores within one's racial group"(Rogelberg 2007) promoted by the U.S. Department of Labor, has been used since 1981 given the assumption (or observation) that raw test scores may overpredict the future performance for racial majorities and underpredict for racial minorities. In 1991, the practice of race norming is considered to be illegal in employment related tests after the passage of the Civil Rights Act of 1991, and eighteen years later in 2009, a similar decision made by the U.S. Supreme Court in *Ricci v. DeStefano*, 557 U.S. 557 (2009) has become a landmark precedent on disparate impact liability. See Appendix A for more discussions on the disparate impact standard.

As discussed in Section 2.2, disparate impact discrimination is the U.S. version of indirect discrimination and intends to cover unintentional discrimination only. In Section 3.1, we illustrated this standard is considered to achieve group-level parity based on a protected attribute, Definition 5 in Section 3 and `Model 3` in Sections 4 and 5 all can satisfy this standard. Broadly speaking, we can think of this standard as a stricter regulation than prohibiting indirect discrimination by prohibiting proxy variables.

6.1.7 Community Rating

At the other extreme, community rating contrasts with risk rating aiming to ensure group fairness on all variables as protected features – everyone pays the same premium on the same insurance product, whereas most of the regulations discussed earlier in this section still allow insurance products to be risk rated (to different extent).

For example, health care is often viewed as a social good, and therefore in some jurisdictions health insurance (or health system) is based on a system of community rating. In Australia, after the introduction of the National Health Act 1953 and the Private Health Insurance Act 2007 by the Australian government, private health insurance is community rated regardless of factors such as health status, age, claims history or pre-existing conditions of individuals (i.e., medical factors for underwriting).

6.1.8 Affirmative Action

An affirmative action practice or policy that seeks to improve the representation of historically excluded groups that were underrepresented and unfairly discriminated against in the past, most commonly in the fields of employment and education. In particular, the practice of

discriminatory insurance practices with regards to homeowners insurance.

affirmative action not only aims to eliminate discrimination or achieve fairness, but also to redress past discrimination and remediate its effects, and hence its practice may give preferential treatment to historically disadvantaged groups, which is also know as (intentional) positive discrimination as the opposite of intentional unfair discrimination (under the scenario of no regulation).

In general, it is difficult to envisage affirmative action can be applied to insurance products.

6.1.9 Other Regulations

Some other more specific or broad regulations that cannot be classified into the regulations discussed earlier in this section are listed below:

1. **Regulatory Prior Approval.** As compared with no regulation, a more realistic minimum standard for insurance companies is that the variables they use need to be approved in advance by regulators. In the United States, a general standard is that insurance rates shall not be excessive, inadequate or unfairly discriminatory, and state regulators usually require insurance companies to prove that all rating factors are actuarially sound by their predicted value of future losses; and then exceptions are made for certain protected variables that require (special) social protection, even if these variables are actuarially justified.

2. **Regulatory Prior Approval Factors.** A more comprehensive approach to prohibiting the use of protected or proxy variables is to provide insurance companies with a list of acceptable factors to choose from. In California, according to Proposition 103, automobile insurers should consider three mandatory rating factors more heavily than other factors in decreasing order of importance: 1) the insured's driving safety record; 2) the number of miles he or she drives annually; 3) the number of years of driving experience the insured has; and in addition insurers are allowed to use 15 optional rating factors including type of vehicle, type of use of vehicle, vehicle characteristics, academic standing and marital status of the rated driver; see California Code of Regulations (10 CCR § 2632.5[24]). As for another more restrictive example, under the ACA, insurers are only allowed to consider insureds' family size, rating area, age and smoking status, and this practice in healthcare system is also known as adjusted community rating since the use of health status, claims experience or gender is not allowed.

3. **Prohibition as a Sole Factor.** Insurers can be prohibited to use a certain factor as the sole basis in underwriting or rating decisions, such as zip code or credit score, and this regulation can be regarded as a special way of restricting protected or proxy variables.

4. **Disparate Impact Standard With Flexibility (E.g., the Four-Fifths Rule).** Corresponding to Definition 7 in Section 3 and `Model 4` in Sections 4 and 5, this regulation is a relaxation of disparate impact standard as it is less rigorous, and if applied to the insurance field, insurance companies will be allowed to legally deviate from group fairness criterion to a certain extent. As a classic example in the employment

[24]See available at https://www.law.cornell.edu/regulations/california/10-CCR-Sec-2632-5.

context, the Four-Fifths Rule is codified in the 1978 *Uniform Guidelines for Employee Selection Procedures*[25] as follows:

> a selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

Currently, this method is often used along with more rigorous statistical tests and has been applied in many other fields as a reference dividing line of disparate impact.

5. **Restriction on the Influence of Protected (or Proxy) Variables.** For example, as discussed above, California's Proposition 103 requires insurance companies[26] to base automobile insurance premiums primarily upon three mandatory rating factors within driver's control.

6. **Effective Prohibition Through Insurance Companies.** For various reasons, some insurance companies may voluntarily not use certain protected variables (e.g., occupation and educational level), and if major insurers do not use or collect a specific protected (or proxy) variable – this variable is effectively prohibited, although it is not regulated by laws or regulations.

6.2 Comparison between Different Regulations

Table 2: Comparison between Different Regulations

Regulation	Fairness Criteria	Representative Model[27]
No Regulation	Neither	Model 1
Restriction on a Protected Variable	Neither	Model 1*
Prohibition on a Protected Variable	Individual	Model 2 or 5
Restriction on a Proxy Variable	Individual	Model 2*
Prohibition on a Proxy Variable	Individual	Model 2*[28]
Disparate Impact Standard	Group	Model 3 or 4
Community Rating	Group	Model 3 or 4
Affirmative Action	Neither	None

[25]See available at https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml

[26]Specifically, insurers should perform a sequential analysis, see https://www.insurance.ca.gov/0250-insurers/0800-rate-filings/upload/Sequential-Analysis.pdf and http://www.insurance.ca.gov/0250-insurers/0800-rate-filings/upload/Class-Plan-Instructions-02_10_2020.pdf.

[27]For representative models, Model 1 denotes the original `Model 1` discussed in Sections 4 and 5, and Model 1* denotes `Model 1` or `Model 1` with some adjustments – for example, if insurers are restricted from charging extremely high premiums for policyholders from a high risk category that exceeds a certain threshold compared to the reference category, we may need to adjust the model coefficients for the high risk category to meet regulatory requirements.

In Section 6.1, all the regulations are sorted according to their strictness – from the least restrictive "no regulation" to the most restrictive "community rating", and in this subsection, we attempt to match these regulations with individual fairness or group fairness, and at the same time list the corresponding model(s) from those discussed in Sections 4 and 5 that directly satisfy these regulations (i.e., models not listed in the last column of Table 2 may still satisfy these regulations), as summarized in Table 2.

From another point of view, "no regulation" is our baseline scenario where insurers can adopt risk-based pricing without restrictions, while all other regulations deviate more or less from risk-based pricing involving subsidies from low-risk individuals to high-risk individuals according to each individual's group membership based on a protected characteristic. Since direct discrimination is allowed under the scenarios of "no regulation" and "restriction on a protected variable", both regulations belong to neither individual fairness nor group fairness. "Prohibition on a protected variable" is equivalent to requiring same premiums and benefits regardless of group membership provided that all other rating factors remain the same, and insurers will face less restrictive regulations and only need to ensure fairness at the individual level; while the "disparate impact standard" aims to achieve fairness at the group level –individuals in the high-risk group pay roughly the same average premium as the low-risk group. At the other extreme, "affirmative action" may intentionally give preferential treatment to historically disadvantaged groups, and if they are also high risk groups for insurers, we will achieve the largest subsidies between groups of all the regulations discussed – positive discrimination occurs.

7 Conclusion

Insurers benefit from the collection of more granular data and the use of more advanced analytics techniques in an age of Big Data, but they are also capable of discriminating against protected classes more efficiently in underwriting or pricing decisions. In this paper, we have established a connection between various insurance regulations, fairness criteria and anti-discrimination insurance pricing models, and in particular we have matched the traditional conflict between individual and group fairness with the opposing views on anti-discrimination regulations between high-risk consumers (or regulators) and insurers, which also reflects the different views of insurance in different contexts: economic commodity or social good (Frees and Huang 2021).

Our empirical analysis using both GLM and XGBoost compares the outcome of different models and analyse their impact from the perspectives of fairness-acuracy trade-off, adverse selection, and solidarity. Overall, Model 1 has the best prediction accuracy and worst group fairness in all scenarios and XGBoost has better prediction accuracy compared to GLM. We also find that Models 2-5 satisfy the four-fifth rule in all scenarios considered in this paper and different models can achieve the best trade-off under different scenarios. We also find that GLM and XGBoost may have different sensitivity to protected variables. For example, XGBoost method is generally more sensitive to small gender-related differences,

[28] As a further extension of `Model 2`, the proxies of a protected variable are also prohibited from being used and therefore removed from modelling.

and we suggest insurance regulators or practitioners need to be aware that different pricing methods may have different degrees of sensitivity to the protected variable. We find that in certain scenarios analysed in the paper, insurers implementing a fair model will lose relatively high-risk customers and gain low-risk customers for both GLM and XGBoost models. And fairness is achieved mostly via the subsidy from young females to young males considering the actual claims cost as the benchmark.

This research contributes to the understanding and mitigation of indirect discrimination in the insurance industry and here we propose some research directions of future work. First, anti-discrimination regulations vary across different lines of business and jurisdictions while our research primarily focuses on general insurance. Other lines of business are worth of further study. Second, one practical difficulty is how to collect protected policyholder information (such as race or ethnicity) and Lindholm et al. (2022) indicate indirect insurance discrimination is caused by incomplete discriminatory information. Recently, several methods in the machine learning field have been proposed to deal with this issue, see Kallus, Mao, and Zhou (2021) and Wang et al. (2020). More research in the insurance domain is needed. Third, there could be other ways to mitigate the impact of insurance discrimination in practice, such as developing assessment tools for regulators or simply controlling the effect of protected features instead of prohibiting their use. The assessment tools could also help make the regulation policies better clarified. Fourth, future studies may further investigate the impacts of new technologies and innovations on insurance discrimination issues, such as telematics on auto insurance (i.e., more positive impacts) or genetic testing on life and health insurance (i.e., potential negative impacts).

Appendices

Appendix A: The Evolvement of U.S. Insurance Discrimination Regulations

On July 6, 2021, Colorado Senate Bill (SB) 21-169[29] was signed into law, and this legislative reform is considered a breakthrough attempt on the issue of indirect insurance discrimination in insurance regulations. In Appendix 1, we summarize the evolvement of U.S. insurance discrimination regulations, including existing insurance discrimination definitions that have been widely used and some newly proposed definitions being considered by insurance regulators from various stakeholders.

Part 1. State Based Insurance Regulation and Unfair Discrimination Statutes

The McCarran-Ferguson Act of 1945 formally delegated the regulatory authority from Congress to the states[30] regarding the regulation of the business of insurance, and therefore general insurers are regulated predominantly at the state level, including our focus – anti-discrimination laws and regulations in the insurance industry.

Wortham (1986a) reviewed the history of the development of state unfair discrimination statutes in relation to insurance discrimination, which often require that insurance classification should be supported by statistical evidence showing correlation with loss (in all states in at least some personal lines of insurance). In particular, state insurance laws often require that insurance premiums are fair, not unfairly discriminatory – a general standard that is commonly contained in insurance regulations that insurance rates shall not be excessive, inadequate or unfairly discriminatory, and it is usually defined as follows, derived from Section 5A (3) Unfairly Discriminatory Rates of the NAIC Property and Casualty Model Rating Law (GDL-1775):

> Unfair discrimination exists if, after allowing for practical limitations, price differentials fail to reflect equitably the differences in expected losses and expenses. A rate is not unfairly discriminatory if it is averaged broadly among persons insured under a group, franchise or blanket policy or a mass marketed plan.

As another example, the Casualty Actuarial Society (CAS) published the Statement of Principles Regarding Property and Casualty Insurance Ratemaking in 1988 which describes in Principle 4 that

> A rate is reasonable and not excessive, inadequate, or unfairly discriminatory if it is an actuarially sound estimate of the expected value of all future costs associated with an individual risk transfer.

[29] SB21-169: Protecting Consumers from Unfair Discrimination in Insurance Practices, see available at https://doi.colorado.gov/for-consumers/sb21-169-protecting-consumers-from-unfair-discrimination-in-insurance-practices and https://leg.colorado.gov/bills/sb21-169.

[30] See the legislative background of the Act as per the NAIC, available at <https://content.naic.org/cipr_topics/topic_mccarran_ferguson_act.htm>

The above principle also corresponds to the actuarial fairness principle – a guiding principle in insurance industry, see Landes (2015) for more discussion.

Part 2. Disparate Impact Standard and Its Applicability in the Insurance Industry

Disparate impact, also known as adverse impact, refers to discrimination that is unintentional, and it is a legal term as a means of proving that indirect discrimination has occurred without the need to prove discriminatory intent or motive in a discrimination lawsuit.

In the landmark ruling of *Griggs v. Duke Power Co.* (1971)[31], the first legal precedent was established for disparate impact claims under Title VII of the Civil Rights Act of 1964 in the employment context. In a disparate impact case for employment discrimination, a **three-step burden-shifting approach** is adopted[32]. First, the plaintiff must establish a prima facie case of adverse disparate impact. It is important to note that even if where a disparate impact is shown by a plaintiff at step one, the practice would not constitute discrimination (or impose liability) if the defendant (i.e., the employer) can demonstrate the practice causing a disparate impact is consistent with business necessity at step two (i.e., a business necessity test constitutes a defence to disparate impact claims). Finally, if the employer has successfully passed the business necessity test for their discriminatory practice, the employee then has the opportunity and may still succeed at step three if they can show that an alternative practice exists that is comparable and less discriminatory, where the employer refuses to adopt. On November 21, 1991, the disparate impact framework (a.k.a., the disparate impact theory of liability) was codified into the Civil Rights Act of 1991 in response to several controversial and adverse U.S. Supreme Court decisions[33] prior to the introduction of the Act. In *Smith v. City of Jackson* (2005)[34], the U.S. Supreme Court confirms that the Age Discrimination in Employment Act of 1967 (ADEA) also authorizes disparate impact claims, but "the scope of disparate-impact liability under ADEA is narrower than under Title VII"[35].

[31]Griggs v. Duke Power Co., 401 U.S. 424 (1971).

[32]Disparate impact discrimination is also applicable under Title VI. According to the U.S. Department of Transportation, "Disparate impact (also called adverse impact) discrimination happens under Title VI when a recipient of federal funds from FHWA adopts a procedure or engages in a practice that has a disproportionate, adverse impact on individuals who are distinguishable based on their race, color, or national origin – even if the recipient did not intend to discriminate", available at https://www.fhwa.dot.gov/civilrights/programs/docs/Title%20VI%20-%20Types%20of%20Discrimination.pdf; and similarly for more detail on the three-step approach regarding how to prove a violation of disparate impact standard under Title VI, see Title VI Legal Manual published by the U.S. Department of Justice (DOJ), available at https://www.justice.gov/crt/fcs/T6Manual7#:~:text=To%20establish%20an%20adverse%20disparate,and%20(4)%20establish%20causation.

[33]Including *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642 (1989); see also Civil Rights Act of 1991 § 2(2), "the decision of the Supreme Court in Wards Cove Packing Co. v. Atonio, 490 U.S. 642 (1989) has weakened the scope and effectiveness of Federal civil rights protections"; the *Wards Cove* case's precedent was nullified by the 1991 Act since this precedent would make it extremely difficult for the plaintiff to prove disparate impact claims under Title VII.

[34]Smith v. City of Jackson, 544 U.S. 228 (2005).

[35]See opinion of the Court in *Smith v. City of Jackson*, 544 U.S. 228 (2005), available at https://www.law.cornell.edu/supct/pdf/03-1160P.ZO.

Since the disparate impact theory was proposed, for a long time, it is generally believed that disparate impact standard was only applicable to the field of employment discrimination[36]. On February 8, 2013, the U.S. Department of Housing and Urban Development (HUD) issued a final rule[37] entitled "Implementation of the Fair Housing Act's Discriminatory Effects Standard" ("the 2013 rule") that authorizes disparate impact claims under the Fair Housing Act (FHA) as a formal interpretation of the Act, consistent with HUD's long-held view. In particular, HUD restated its position that the Fair Housing Act applies to homeowners insurance[38], and hence **disparate impact standard is applicable to prohibit discriminatory insurance practices with regards to homeowners insurance**.

However, the application of disparate impact standard in the context of the insurance industry is strongly opposed by insurance companies. The National Association of Mutual Insurance Companies (NAMIC) claimed that the 2013 rule meant to insurers that "any factor used by insurers to assess risk could be challenged if it produced statistically disproportionate outcomes among demographic groups"[39], and argued that "insurers do not even know the race, religion, or national origin of their insureds"[40]. Therefore, NAMIC and American Insurance Association – two insurance industry trade associations jointly challenged the validity of HUD's disparate impact rule under the Fair Housing Act in the U.S. District Court for the District of Columbia, see *American Insurance Association, et al. v. United States Department of Housing and Urban Development, et al.*[41], and scored an initial victory on November 3, 2014.

On June 25, 2015, the U.S. Supreme Court held that disparate impact claims are cognizable under the Fair Housing Act in the landmark decision of *Texas Department of Housing and*

[36]In particular, in Alexander v. Sandoval, 532 U.S. 275 (2001), U.S. Supreme Court Held Title VI statute does not allow for private lawsuits based on disparate impact, see https://www.fhwa.dot.gov/civilrights/programs/docs/Title%20VI%20-%20Intentional%20Discrimination%20and%20Disparate%20Impact.pdf.

[37]See the 2013 rule issued by HUD on February 8, 2013, "Implementation of the Fair Housing Act's Discriminatory Effects Standard", available at https://www.hud.gov/sites/documents/DISCRIMINATORYEFFECTRULE.PDF.

[38]See the 2013 rule, HUD responded to the concerns from the insurance industry that "HUD has long interpreted the Fair Housing Act to prohibit discriminatory practices in connection with homeowner's insurance, and courts have agreed with HUD, including in *Ojo v. Farmers Group*. Moreover, as discussed above, HUD has consistently interpreted the Act to permit violations to be established by proof of discriminatory effect. By formalizing the discriminatory effects standard, the rule will not, as one commenter suggested, 'undermine the states' regulation of insurance.'... McCarran-Ferguson does not preclude HUD from issuing regulations that may apply to insurance policies."

[39]See NAMIC, Our Positions – Disparate Impact Rule, available at https://www.namic.org/issues/disparate-impact-rule;

[40]See NAMIC's letter to the NAIC: NAMIC comments on the draft NAIC Principles on Artificial Intelligence, available at https://content.naic.org/sites/default/files/call_materials/NAMIC%20-%20NAIC%20AIWG%20-%20Comments%20-%206-29-20.pdf.

[41]Civil Case No. 13–00966 (RJL), United States District Court, District of Columbia. Signed November 7, 2014, see https://ecf.dcd.uscourts.gov/cgi-bin/show_public_doc?2013cv0966-47; see also <https://www.lawyerscommittee.org/project/aianamic/>, "Judge Leon, accepting plaintiffs' argument that the FHA only prohibits intentional discrimination and that the McCarran-Ferguson Act forecloses the application of disparate impact theory to the provision of homeowners' insurance, held the FHA unambiguously forecloses the possibility of disparate impact claims."

Community Affairs v. Inclusive Communities Project, Inc. (2015)[42], and the NAMIC's victory at the District Court level was overruled by the Supreme Court decision in *Inclusive Communities*. In April 2016, the plaintiffs filled an amended complaint to challenge the HUD's disparate impact rule and a summary judgment motion was filed in June 2016 that seeks to "invalidate the 2013 Rule to the extent it applies to insurers' ratemaking and underwriting decisions."(Willis, Andreano, and Sommerfield 2021) and the lawsuit is currently pending in the D.C. federal district court since June 2018 for HUD's revisions to the 2013 rule.

On September 24, 2020, HUD issued a final rule[43] entitled "HUD's Implementation of the Fair Housing Act's Disparate Impact Standard" ("the 2020 rule") that amended the 2013 rule, including the "clarification regarding the application of the standard to state laws governing the business of insurance." Since being proposed, the 2020 rule has been widely criticized by consumer advocates and Democratic lawmakers because it requires heavy burden of proof on the plaintiff for a disparate impact claim under the Fair Housing Act[44] and appears to favour the defendant, and therefore there are at least three lawsuits challenging the 2020 rule in federal district courts. On January 26, 2021, President Biden issued an executive order[45] to direct HUD to review the effects of the 2020 rule, including the effect that revising the 2013 rule. On 25 June, 2021, HUD formally proposes to rescind the 2020 rule and restore the 2013 rule[46].

In terms of the definition, the United States does not clearly define disparate impact (even indirect discrimination) in statute law. In short, disparate impact describes "when a facially neutral practice that has an unjustified adverse impact on members of a protected class." In other fields, for example with regard to fair lending, the Federal Reserve interprets disparate treatment and disparate impact with respect to lending discrimination under the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act[47] as follows:

> **Disparate Treatment:** the existence of illegal disparate treatment may be established either by statements revealing that a lender explicitly considered prohibited factors (overt evidence) or by differences in treatment that are not fully explained by legitimate nondiscriminatory factors (comparative evidence).
>
> **Disparate Impact:** a disparate impact occurs when a lender applies a racially

[42]Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. 519 (2015).

[43]See the 2020 rule issued by HUD on September 24, 2020, "HUD's Implementation of the Fair Housing Act's Disparate Impact Standard", available at https://www.govinfo.gov/content/pkg/FR-2020-09-24/pdf/2020-19887.pdf.

[44]See a comprehensive summary for the differences in the HUD's 2020 Rule, and how the *Inclusive Communities* decision in 2015 is different from the HUD's 2013 Rule: https://www.jdsupra.com/legalnews/hud-issues-final-rule-on-the-fair-63161/.

[45]See the executive order available at https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/26/memorandum-on-redressing-our-nations-and-the-federal-governments-history-of-discriminatory-housing-practices-and-policies/.

[46]On 25 June, 2021, HUD published the Proposed Rule entitled "Reinstatement of HUD's Discriminatory Effects Standard," see available at https://www.federalregister.gov/documents/2021/06/25/2021-13240/reinstatement-of-huds-discriminatory-effects-standard.

[47]See the Federal Reserve, "Federal Fair Lending Regulations and Statutes Overview", available at https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_over.pdf;

> (or otherwise) neutral policy or practice equally to all credit applicants but the policy or practice disproportionately excludes or burdens certain persons on a prohibited basis. . . . Although the law on disparate impact as it applies to lending discrimination continues to develop, it has been clearly established that a policy or practice that creates a disparity on a prohibited basis is not, by itself, proof of a violation.

To sum up, we believe the definition of disparate treatment includes direct discrimination and some intentional indirect discrimination and disparate impact is a subset of indirect discrimination because it is a legal definition and only intends to cover unintentional discrimination (although it may cover intentional indirect discrimination that is too difficult to prove discriminatory intent under a disparate treatment case). For example, redlining as a classic example of indirect discrimination (or proxy discrimination) is a form of (illegal) disparate treatment[48], rather than disparate impact.

Part 3. Colorado Bill and Recent Regulatory Reform Discussion

On August 14, 2020, the NAIC published guiding principles on artificial intelligence (AI)[49] including a key principle "encouraging industry participants to take proactive steps to avoid proxy discrimination against protected classes when using AI platforms[50]" developed by the NAIC's Big Data and Artificial Intelligence Working Group. Specifically, as part of the "fair and ethical" tenet, one key NAIC's AI principle is outlined below:

> Consistent with the risk-based foundation of insurance, AI actors should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for consumers and to **avoid proxy discrimination** against protected classes. AI systems should not be designed to harm or deceive people and should be implemented in a manner that **avoids harmful or unintended consequences** and corrects and remediates for such consequences when they occur.

However, the term "proxy discrimination" has not yet been defined by the NAIC and it is unclear for insurers on how to comply with the guiding principles to avoid proxy discrimination in practice. In addition, the guiding principle also covers "unintended consequences", which could upend the industry's understanding of indirect discrimination, whereas insurance anti-discrimination laws were previously thought to have generally focused on direct discrimination and intentional indirect discrimination.

[48] See the Federal Reserve, available at https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_over.pdf; "Redlining is a form of illegal disparate treatment whereby a lender provides unequal access to credit, or unequal terms of credit, because of the race, color, national origin, or other prohibited characteristic(s) of the residents of the area in which the credit seeker resides or will reside or in which the residential property to be mortgaged is located. Redlining may violate both the FHAct and the ECOA."

[49] See National Association of Insurance Commissioners (NAIC) Principles on Artificial Intelligence (AI), available at https://content.naic.org/sites/default/files/inline-files/AI%20principles%20as%20Adopted%20by%20the%20TF_0807.pdf, as a response to the OECD Principles on Artificial Intelligence.

[50] See NAIC Unanimously Adopts Artificial Intelligence Guiding Principles, available at https://content.naic.org/article/news_release_naic_unanimously_adopts_artificial_intelligence_guiding_principles.htm#:~:text=Washington%20(August%2020%2C%202020),safe%2C%20secure%20and%20robust%20outputs.

A recent example is the Colorado Senate Bill 21-169 in the United States, which was passed and signed into law in July 2021, and its definition of unfair discrimination has a "disparate impact" component, which is outlined as follows

> "Unfairly discriminate" and "unfair discrimination" include the use of one or more external consumer data and information sources, as well as algorithms or predictive models using external consumer data and information sources, that have a correlation to race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression, and that use **results in a disproportionately negative outcome** for such classification or classifications, which negative outcome **exceeds the reasonable correlation** to the underlying insurance practice, including losses and costs for underwriting.

The above definition could be the first insurance regulation to focus on the effects of discrimination at the group level – that is common in other areas such as lending, housing or college admissions.

Appendix B: Implementation Details of Generalized Linear Models (GLMs) and Extreme Gradient Boosting (XGBoost)

Generalized Linear Models (GLMs)

Generalized linear models (GLMs) have been widely used by actuaries in general insurance pricing. In this paper, we adopt the classical frequency-severity approach by building two separate frequency and severity models. Following De Jong and Heller (2008) and Frees, Derrig, and Meyers (2014), the structure of GLMs is as follows:

$$g(\mu) = x'\beta$$

let Y denote the response variable, x the explanatory variables and μ the expectation of Y, where its distribution is a member of the exponential family of distributions, and $g(\cdot)$ denote the monotonic link function. To illustrate frequency and severity models, let N denote the number of claims, E the exposures and S the (aggregate) claim amount. For a claim frequency model, we adopt the Poisson regression model, which is typically applied for count data, with an offset term for exposures:

$$N \sim \text{Poisson}(\lambda), \quad ln(\lambda) = ln(E) + x_F'\beta_F$$

where x_F is the set of covariates used in modeling frequency and β_F is the corresponding set of regression coefficients. The claim count N follows a Poisson distribution, and the log link is chosen $g(\cdot) = ln(\cdot)$ and logarithmic exposure is added as an offset variable accounting for the premium is proportional to the exposures. For a claim severity model, we apply the gamma regression model for claim size conditional on the event that there is at least one claim filed by policyholders (i.e., $N > 0$):

$$S/N \sim \text{Gamma}(\alpha, \gamma), \quad ln(\alpha) = x_S'\beta_S$$

where x_S is the set of covariates used in modeling severity and β_S is the corresponding set of regression coefficients, and the sets of covariates are not necessarily the same for the frequency and severity models . The claim size S/N follows a gamma distribution, also with a logarithmic link function $g(\cdot) = ln(\cdot)$.

Extreme Gradient Boosting (XGBoost)

XGBoost was proposed by Chen and Guestrin (2016) as a novel gradient tree boosting method and has rapidly gained popularity due to its high efficiency in computational speed and predictive performance in applications in many fields, including its superior performance in machine learning competitions on Kaggle.

XGBoost requires to build classification and regression trees (or CARTs) iteratively, where each subsequent tree (also known as a weaker learner) is trained to predict the residuals of the previous tree, so that by building more weak learners sequentially, the training continuous and each new tree corrects errors in the previous tree until a stopping criterion is reached. In the XGBoost, assuming we have K additive functions (or trees), a tree ensemble model using K trees is expressed as

$$\hat{Y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

where $\mathcal{F}$ is the space of functions containing all regression trees and each f_k represents an independent tree structure. The XGBoost method minimizes a regularized objective function as follows (Chen and Guestrin 2016) :

$$\mathcal{L} = \sum_{i=1}^{n} l(Y_i, \hat{Y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

$$\text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2$$

where l is a differentiable convex loss function that measures the difference between the observed outcome Y_i and predicted outcome $\hat{Y}_i$, and $\Omega(f)$ is the regularization term that penalizes the complexity of the model, where T is the number of leaves in the tree and w is the sum of leaf weights, and γ and λ are the regularization hyperparameters. In addition, the training of the XGBoost model is additive (Chen and Guestrin 2016) and at the t-th iteration, the objective we aim to minimize is expressed as follows:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(Y_i, \hat{Y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

where $\hat{Y}_i^{(t-1)}$ is the prediction at the $(t-1)$-th iteration. For more mathematical details on implementing the XGBoost method, we refer interested readers to Chen and Guestrin (2016). To apply XGBoost to insurance pricing, the loss functions for claim frequency and severity models need to be appropriately specified. In fact, the choice of learning objective function in XGBoost is similar to the choice of the distributions of Y in GLMs, we set the learning objective to `count:poisson` (Poisson regression) for claim count and `reg:gamma` (gamma

regression with log-link) for claim size. For all XGBoost models, we perform a grid search for tuning hyperparameters by steps using five-fold cross-validation, and we refer interested readers to Fauzan and Murfi (2018) for detailed grid search scheme.

Appendix C: Supplementary Figures to Section 5

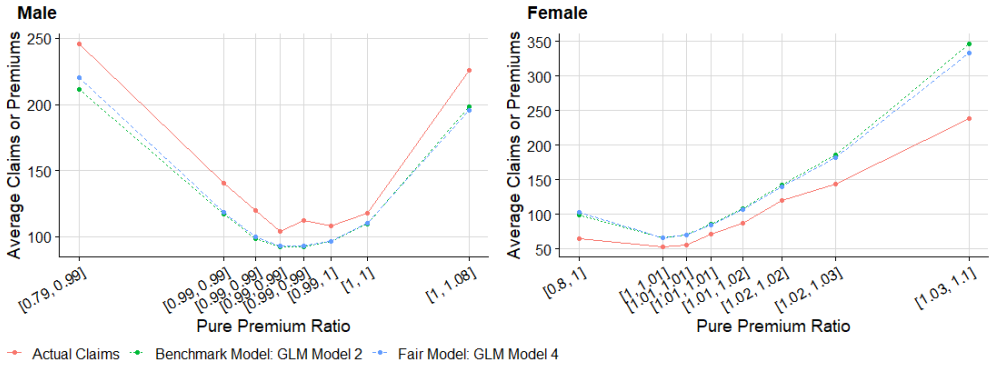

Figure 11: Double Lift Charts By Gender (GLM Model 4 vs. GLM Model 2)

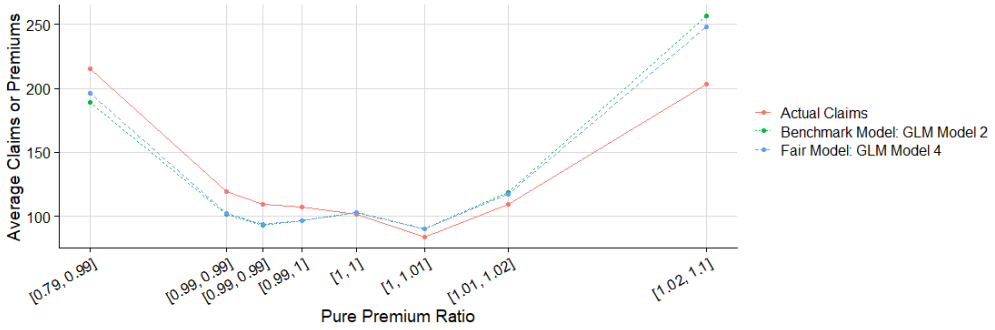

Figure 12: Double Lift Chart (GLM Model 4 vs. GLM Model 2)

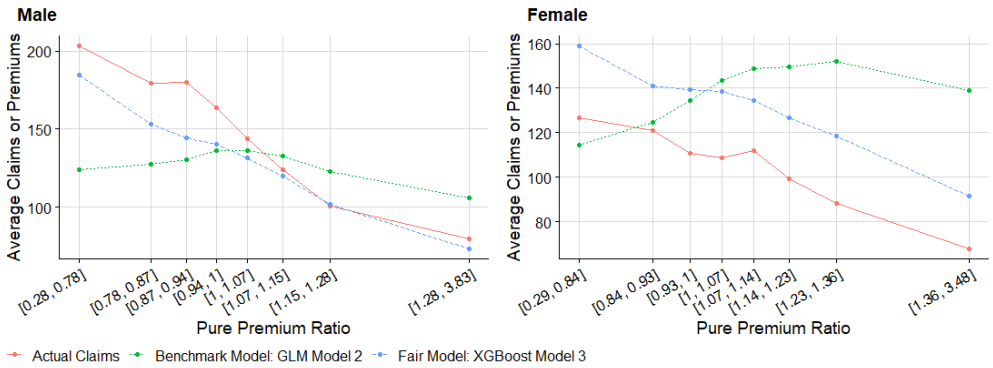

Figure 13: Double Lift Charts By Gender (XGBoost Model 3 vs. GLM Model 2)

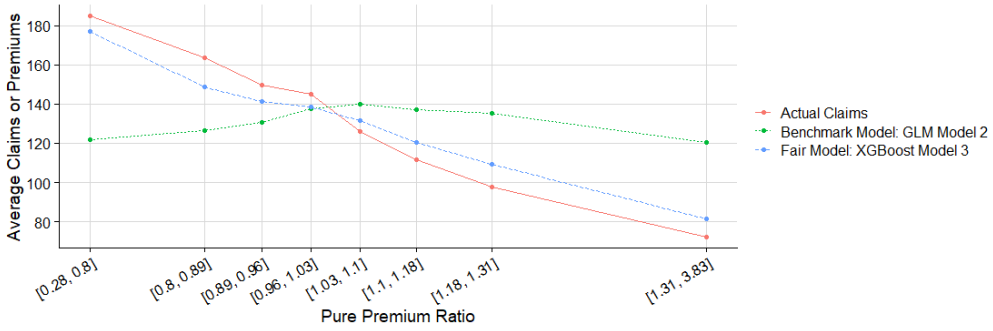

Figure 14: Double Lift Chart (XGBoost Model 3 vs. GLM Model 2)

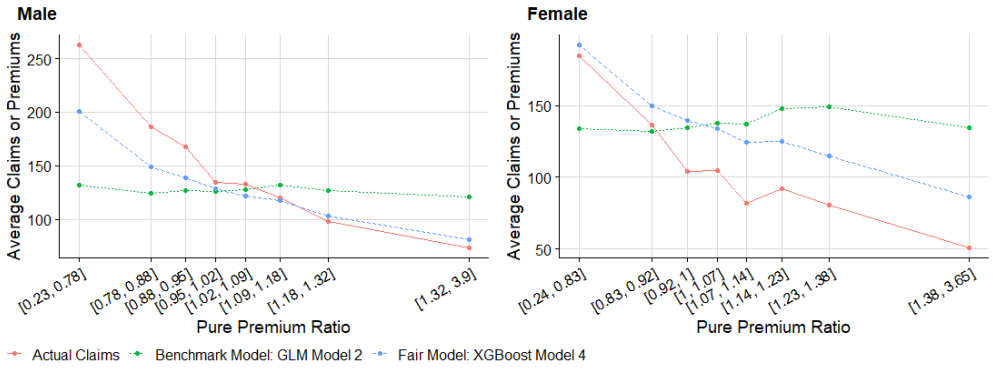

Figure 15: Double Lift Charts By Gender (XGBoost Model 4 vs. GLM Model 2)

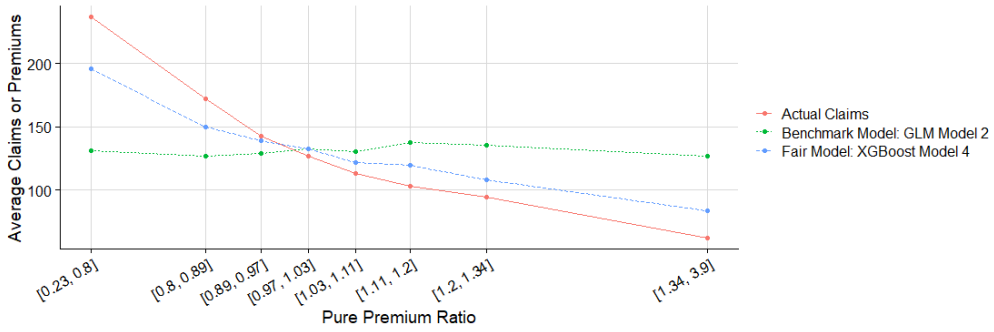

Figure 16: Double Lift Chart (XGBoost Model 4 vs. GLM Model 2)

Figure 17: Relative and Average Premium Difference (XGBoost Models vs. GLM Model 2)

Figure 18: Relative and Average Premium Difference (XGBoost Models vs. GLM Model 1)

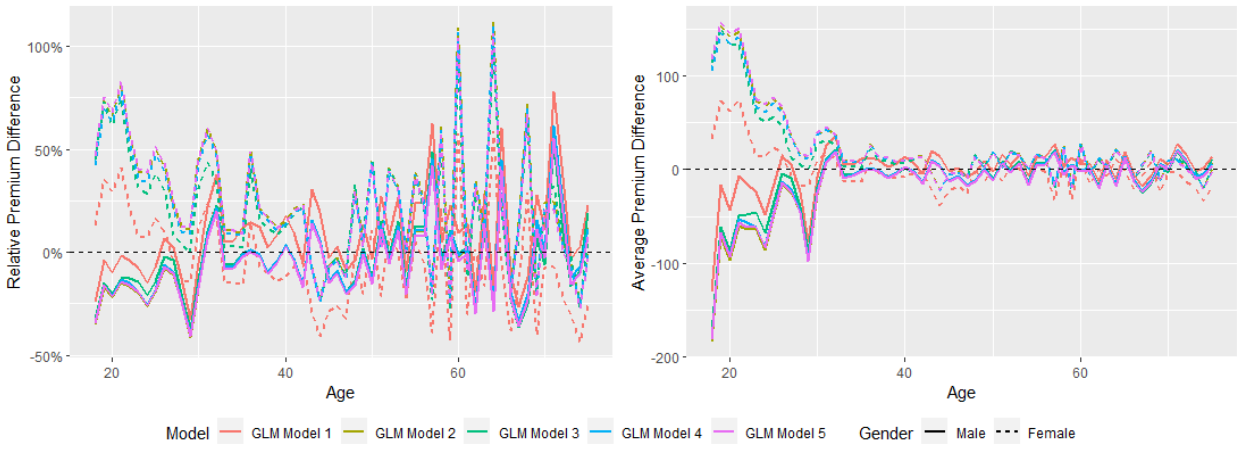

Figure 19: Relative and Average Premium Difference (GLM Models vs. Actual Claim Costs)

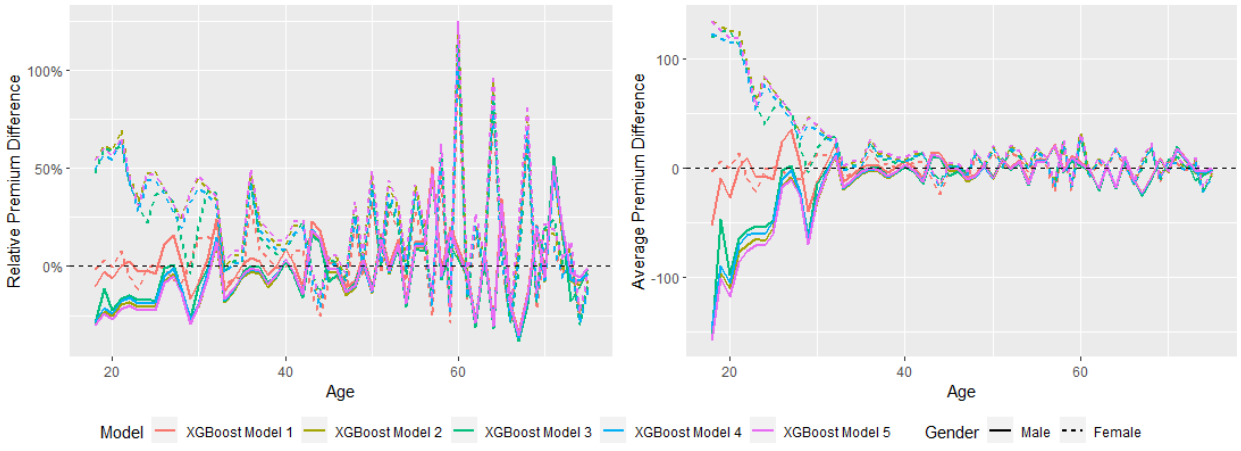

Figure 20: Relative and Average Premium Difference (XGBoost Models vs. Actual Claim Costs)

References

Agarwal, Alekh, Miroslav Dudik, and Zhiwei Steven Wu. 2019. "Fair Regression: Quantitative Definitions and Reduction-Based Algorithms." In *International Conference on Machine Learning*, 120–29. PMLR.

AHRC. 2020. "Using Artificial Intelligence to Make Decisions: Addressing the Problem of Algorithmic Bias." https://humanrights.gov.au/sites/default/files/document/publication/ahrc_technical_paper_algorithmic_bias_2020.pdf.

Australian Law Reform Commission. 2003. *Essentially Yours–the Protection of Human Genetic Information in Australia, Volume 1 and Volume 2. Report 96.* /newline/url%7Bhttps://www.alrc.gov.au/publication/essentially-yours-the-protection-of-human-genetic-information-in-australia-alrc-report-96/%7D.

Avraham, Ronen, Kyle D Logue, and Daniel Schwarcz. 2014a. "Towards a Universal Framework for Insurance Anti-Discrimination Laws." *Conn. Ins. LJ* 21: 1.

———. 2014b. "Understanding Insurance Antidiscrimination Laws." *Southern California Law Review* 87 (2): 195–274.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

Barocas, Solon, and Andrew D Selbst. 2016. "Big Data's Disparate Impact." *Calif. L. Rev.* 104: 671.

Berk, Richard. 2009. "The Role of Race in Forecasts of Violent Crime." *Race and Social Problems* 1 (4): 231–42.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. "A Convex Framework for Fair Regression." *arXiv Preprint arXiv:1706.02409*.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research*, 0049124118782533.

Binns, Reuben. 2018. "Fairness in Machine Learning: Lessons from Political Philosophy." In *Conference on Fairness, Accountability and Transparency*, 149–59. PMLR.

———. 2020. "On the Apparent Conflict Between Individual and Group Fairness." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–24.

Birnbaum, Birny. 2020. "Insurance Consumer Protection Issues Resulting from, or Heightened by Covid-19: Presentation to Naic Consumer Liaisoncommittee." https://content.naic.org/sites/default/files/call_materials/Slideshow_Consumer%5B1%5D.pdf.

Caton, Simon, and Christian Haas. 2020. "Fairness in Machine Learning: A Survey." *arXiv Preprint arXiv:2010.04053*.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94.

Chiappa, Silvia. 2019. "Path-Specific Counterfactual Fairness." In *Proceedings of the Aaai Conference on Artificial Intelligence*, 33:7801–8. 01.

Chibanda, Kudakwashe F. 2022. "Defining Discrimination in Insurance." *Cas Research Paper: A Special Series on Race and Insurance Pricing*.

Consumer Reports. 2021. "Effects of Varying Education Level and Job Status on Online Auto Insurance Price Quotes." https://advocacy.consumerreports.org/wp-content/uploads/2021/01/Auto-Insurance-White-Paper-Report-FINAL1.26C.pdf.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." In *Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 797–806.

De Jong, Piet, and Gillian Z Heller. 2008. "Generalized Linear Models for Insurance Data." *Cambridge Books*.

Di Stefano, Pietro G, James M Hickey, and Vlasios Vasileiou. 2020. "Counterfactual Fairness: Removing Direct Effects Through Regularization." *arXiv Preprint arXiv:2002.10774*.

Dolman, Chris, and Dimitri Semenovich. 2019. "Algorithmic Fairness: Some Practical Considerations for Actuaries." *Actuaries Summit 2019*.

Dutang, Christophe, Arthur Charpentier, and Maintainer Christophe Dutang. 2015. "Package 'Casdatasets'."

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness Through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–26.

EIOPA. 2019. *Big Data Analytics in Motor and Health Insurance: A Thematic Review*. Publications Office of the European Union. /newline/url%7Bhttps://register.eiopa.europa.eu/Publications/EIOPA_BigDataAnalytics_ThematicReview_April2019.pdf%7D.

European Commission. 2014. "COMMISSION Staff Working Document Annexes to the Joint Report on the Application of the Racial Equality Directive (2000/43/Ec) and the Employment Equality Directive (2000/78/Ec)." https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52014SC0005.

Fang, Hanming, and Ami Ko. 2018. "Partial Rating Area Offering in the Aca Marketplaces: Facts, Theory and Evidence." National Bureau of Economic Research.

Fauzan, Muhammad Arief, and Hendri Murfi. 2018. "The Accuracy of Xgboost for Insurance Claim Prediction." *Int. J. Adv. Soft Comput. Appl* 10 (2).

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. "Certifying and Removing Disparate Impact." In *Proceedings of the 21th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 259–68.

Frees, Edward W, Richard A Derrig, and Glenn Meyers. 2014. *Predictive Modeling Applications in Actuarial Science*. Vol. 1. Cambridge University Press.

Frees, Edward W, and Fei Huang. 2021. "The Discriminating (Pricing) Actuary." *North American Actuarial Journal, Forthcoming*.

Gaulding, Jill. 1994. "Race Sex and Genetic Discrimination in Insurance: What's Fair." *Cornell L. Rev.* 80: 1646.

Goldburd, Mark, Anand Khare, Dan Tevet, and Dmitriy Guller. 2016. "Generalized Linear Models for Insurance Rating." *Casualty Actuarial Society, CAS Monographs Series* 5.

Hardt, Moritz. 2013. "Fairness Through Awareness." https://course.ece.cmu.edu/~ece734/fall2013/lectures/cmu13-fairness.pdf.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *arXiv Preprint arXiv:1610.02413*.

Henckaerts, Roel, Marie-Pier Côté, Katrien Antonio, and Roel Verbelen. 2021. "Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods." *North American Actuarial Journal* 25 (2): 255–85.

Hutchinson, Ben, and Margaret Mitchell. 2019. "50 Years of Test (Un) Fairness: Lessons for Machine Learning." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58.

III. 2019. "Background on: Credit Scoring." https://www.iii.org/article/background-on-credit-scoring.

Johndrow, James E, and Kristian Lum. 2019. "An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction." *The Annals of Applied Statistics* 13 (1): 189–220.

Kallus, Nathan, Xiaojie Mao, and Angela Zhou. 2021. "Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination." *Management Science*.

Kamiran, Faisal, and Toon Calders. 2012. "Data Preprocessing Techniques for Classification Without Discrimination." *Knowledge and Information Systems* 33 (1): 1–33.

Kasirzadeh, Atoosa, and Andrew Smart. 2021. "The Use and Misuse of Counterfactuals in Ethical Machine Learning." In *Proceedings of the 2021 Acm Conference on Fairness, Accountability, and Transparency*, 228–36.

Kilbertus, Niki, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. "Avoiding Discrimination Through Causal Reasoning." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 656–66.

Kim, Michael P, Omer Reingold, and Guy N Rothblum. 2018. "Fairness Through Computationally-Bounded Awareness." *arXiv Preprint arXiv:1803.03239*.

Kusner, Matt J, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. "Counterfactual Fairness." In *Advances in Neural Information Processing Systems*, 4066–76.

Landes, Xavier. 2015. "How Fair Is Actuarial Fairness?" *Journal of Business Ethics* 128 (3): 519–33.

Lehtonen, Turo-Kimmo, and Jyri Liukko. 2011. "The Forms and Limits of Insurance Solidarity." *Journal of Business Ethics* 103 (1): 33–44.

Lindholm, Mathias, Ronald Richman, Andreas Tsanakas, and Mario V Wüthrich. 2022. "Discrimination-Free Insurance Pricing." *ASTIN Bulletin: The Journal of the IAA* 52 (1): 55–89.

Loi, Michele, and Markus Christen. 2021. "Choosing How to Discriminate: Navigating Ethical Trade-Offs in Fair Algorithmic Design for the Insurance Sector." *Philosophy & Technology*, 1–26.

Meyers, Gert, and Ine Van Hoyweghen. 2018. "Enacting Actuarial Fairness in Insurance: From Fair Discrimination to Behaviour-Based Fairness." *Science as Culture* 27 (4): 413–38.

Miller, Michael J. 2009. "Disparate Impact and Unfairly Discriminatory Insurance Rates." In *Casualty Actuarial Society E-Forum, Winter 2009*, 276.

NAMIC. 2020. "Re: NAMIC Comments on the Draft Naic Principles on Artificial Intelligence." https://content.naic.org/sites/default/files/call_materials/NAMIC%20-%20NAIC%20AI WG%20-%20Comments%20-%206-29-20.pdf.

Pearl, Judea, and others. 2000. "Models, Reasoning and Inference." *Cambridge, UK: CambridgeUniversityPress* 19.

Pope, Devin G, and Justin R Sydnor. 2011. "Implementing Anti-Discrimination Policies in Statistical Profiling Models." *American Economic Journal: Economic Policy* 3 (3): 206–31.

Prince, Anya ER, and Daniel Schwarcz. 2019. "Proxy Discrimination in the Age of Artificial Intelligence and Big Data." *Iowa L. Rev.* 105: 1257.

Rogelberg, Steven G. 2007. *Encyclopedia of Industrial and Organizational Psychology*. Vol. 1. Sage.

Sawyer, Richard L, Nancy S Cole, and James WL Cole. 1976. "Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection." *Journal of Educational Measurement*, 59–76.

Schelldorfer, Jürg, and Mario V Wuthrich. 2019. "Nesting Classical Actuarial Models into Neural Networks." *Available at SSRN 3320525*.

Thorndike, Robert L. 1971. "Concepts of Culture-Fairness." *Journal of Educational Measurement* 8 (2): 63–70.

University of California. 2008. "Race, Sex and Disparate Impact: Legal and Policy Considerations Regarding University of California Admissions and Scholarships." https://regents.universityofcalifornia.edu/regmeet/may08/e2attach.pdf.

Verma, Sahil, and Julia Rubin. 2018. "Fairness Definitions Explained." In *2018 Ieee/Acm International Workshop on Software Fairness (Fairware)*, 1–7. IEEE.

Vincent, Grari, Charpentier Arthur, Lamprier Sylvain, and Detyniecki Marcin. 2022. "A Fair Pricing Model via Adversarial Learning." *arXiv Preprint arXiv:2202.12008*.

Wang, Serena, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael I Jordan. 2020. "Robust Optimization for Fairness with Noisy Protected Groups." *arXiv Preprint arXiv:2002.09343*.

Willis, Christopher J, Richard J Andreano, and Lori Sommerfield. 2021. "President Biden Issues Executive Order Directing Hud to Review Fair Housing Act Disparate Impact Rule." https://www.consumerfinancemonitor.com/2021/02/03/president-biden-issues-executive-order-directing-hud-to-review-fair-housing-act-disparate-impact-rule/.

Wortham, Leah. 1986a. "Insurance Classification: Too Important to Be Left to the Actuaries." *University of Michigan Journal of Law Reform* 19 (2): 349–424.

———. 1986b. "The Economics of Insurance Classification: The Sound of One Invisible Hand Clapping." *Ohio St. LJ* 47: 835.

Wu, Yongkai, Lu Zhang, and Xintao Wu. 2019. "Counterfactual Fairness: Unidentification, Bound and Algorithm." In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Xenidis, Raphaële, and Linda Senden. 2019. "EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination." *Raphaële Xenidis and Linda Senden,'EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination'in Ulf Bernitz et Al (Eds), General Principles of EU Law and the EU Digital Order (Kluwer Law International, 2020)*, 151–82.

Ye, Chenglong, Lin Zhang, Mingxuan Han, Yanjia Yu, Bingxin Zhao, and Yuhong Yang. 2018. "Combining Predictions of Auto Insurance Claims." *arXiv Preprint arXiv:1808.08982*.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. "Learning Fair Representations." In *International Conference on Machine Learning*, 325–33. PMLR.