CAS RESEARCH PAPER
SERIES ON RACE AND INSURANCE PRICING

METHODS FOR QUANTIFYING DISCRIMINATORY EFFECTS ON PROTECTED CLASSES IN INSURANCE

Roosevelt Mosley, FCAS, and Radost Wenman, FCAS

CASUALTY ACTUARIAL SOCIETY

Caveat and Disclaimer
This research paper is published by the Casualty Actuarial Society (CAS) and contains information from various sources. The study is for informational purposes only and should not be construed as professional or financial advice. The CAS does not recommend or endorse any particular use of the information provided in this study. The CAS makes no warranty, express or implied, or representation whatsoever and assumes no liability in connection with the use or misuse of this study. The views expressed here are the views of the authors and not necessarily the views of their current or former employers.

Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance

By Roosevelt Mosley, FCAS, and Radost Wenman, FCAS

Executive Summary

This research paper's main objective is to inspire and generate discussions about algorithmic bias across all areas of insurance and to encourage actuaries to be involved. Evaluating financial risk involves the creation of functions that consider myriad characteristics of the insured. Companies utilize diverse statistical methods and techniques, from relatively simple regression to complex and opaque machine learning algorithms. It has been alleged that the predictions produced by these mathematical algorithms have discriminatory effects against certain groups of society, known as protected classes.

The notion of discriminatory effects describes the disproportionately adverse effect algorithms and models could have on protected groups in society. As a result of the potential for discriminatory effects, the analytical processes followed by financial institutions for decision making have come under greater scrutiny by legislators, regulators, and consumer advocates. Interested parties want to know how to quantify such effects and potentially how to repair such systems if discriminatory effects have been detected.

This paper provides:

- A historical perspective of unfair discrimination in society and its impact on property and casualty insurance.
- Specific examples of allegations of bias in insurance and how the various stakeholders, including regulators, legislators, consumer groups and insurance companies have reacted and responded to these allegations.
- Some specific definitions of unfair discrimination and that are interpreted in the context of insurance predictive models.
- A high-level description of some of the more common statistical metrics for bias detection that have been recently developed by the machine learning community, as well as a brief account of some machine learning algorithms that can help with mitigating bias in models.

This paper also presents a concrete example of an insurance pricing GLM model developed on anonymized French private passenger automobile data, which demonstrates how discriminatory effects can be measured and mitigated.

Introduction

The calls for social justice in the United States have been louder than ever since the events of 2020. While these calls initially were focused on reform in policing, they quickly expanded to address systemic racism in our society. These calls resulted in increased education on the impacts of systemic racism on minorities in the United States, demands for increased equity and the eradication of racism once and for all. The reach of these calls has been wide, and has resulted in corporations, government entities, educational institutions and even religious institutions beginning to examine their businesses and processes to determine how to address racism and social injustice.

The insurance industry has not been exempt from these calls, and this has led to the investigation of the potential impact of systemic racism on insurance underwriting, rating and claims practices. Insurance regulators, consumer advocates and federal and state legislators have held hearings, introduced bills and gathered information on whether insurance rates and/or practices are unfairly discriminatory to protected classes, including minority and low-income policyholders. Because of the questions surrounding insurance rating and the use of models and big data generally, actuaries have played a role in these discussions and will have a role to play in the solutions ultimately implemented. However, if you review actuarial literature, the treatment of discriminatory effects on protected classes in insurance rates is limited.

This research brief will cover the following topics:

- **Section 1: Background** — conditions of unfair discrimination in society and how it has impacted the property and casualty insurance industry
- **Section 2: Accusations of Bias in Insurance** — description of challenges being made to insurance rating, underwriting and claims practices
- **Section 3: What is Unfairly Discriminatory?** — definition of unfair discrimination, including statutory, regulatory and actuarial guidance on unfair discrimination in insurance
- **Section 4: Approaches Measuring and Mitigating Discriminatory Effects on Protected Classes** — explanation of data science methods that have been developed for measuring and controlling bias in models, and how these methods can be applied to actuarial and other insurance predictive models

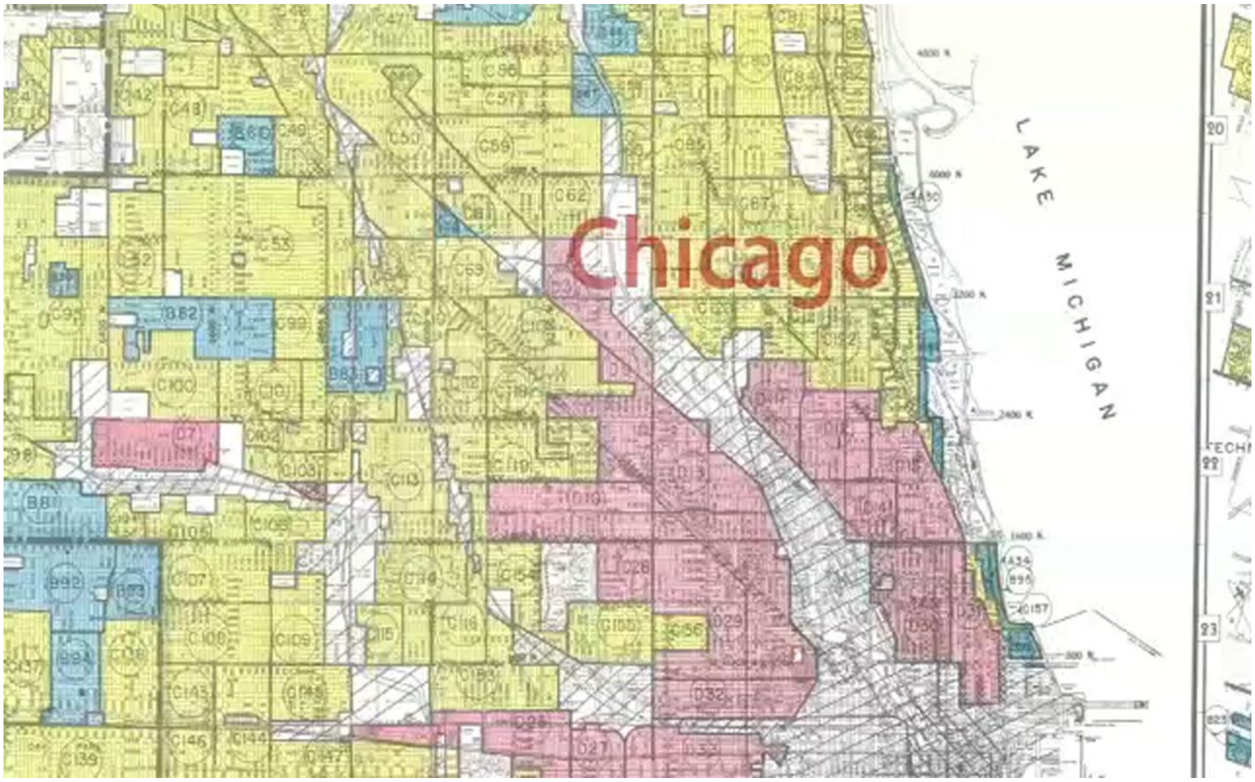
1. Background

Insurance is a key component in building, maintaining and protecting wealth. Financial instruments such as life insurance and annuities allow for building long-term wealth and providing protection for families in the case of unexpected or untimely death. Homeowners and renters insurance provide policyholders with the ability to deploy assets in other ways rather than just saving money in case something happens to their property or possessions. Private passenger automobile insurance enables policyholders to operate vehicles with the knowledge that indemnification is available for accidents for which they may be held liable. Without insurance, our society would not function in an efficient manner and individuals would not be able to build wealth as efficiently or be secure in owning assets of significant value. Actuaries serve key roles in the insurance process as they are often responsible for determining rates that are adequate to cover claims and analyzing reserves and surplus to ensure that companies will be financially viable to pay claims when incurred.

Just as access to insurance provides many benefits, not having access to insurance limits the ability of individuals to build wealth, inhibits the ability of individuals to acquire assets like homes, and does not allow for the safe and confident operation of assets like automobiles. There are examples throughout history when broader issues of racism and discrimination have affected insurance. One example of this in the United States is redlining, a practice which took place during a time when minorities and low-income communities had were effectively denied to financial services available to predominately white communities.

According to the *Encyclopedia of Chicago*, "redlining is the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or poor." Discrimination in financial services existed before the 1930s, but with the creation in 1933 of the Home Owners' Loan Corporation (HOLC), part of President Franklin D. Roosevelt's New Deal, redlining was formally instituted as a policy. Color coded maps were developed to identify levels of risk in lending and insurance. White, affluent areas were categorized as low risk, while minority and poor areas were categorized as high risk and were often highlighted by red lines. Banks and insurance companies adopted the HOLC maps to guide lending and underwriting decisions, and the newly created Federal Housing Administration (FHA) also used the HOLC maps to determine where federally insured new housing construction would take place. See Figure 1 for an example of a HOLC map for the city of Chicago.

Figure 1. HOLC Map of Redlining in Chicago



Source*: Encyclopedia of Chicago*, http://www.encyclopedia.chicagohistory.org/pages/1050.html.

The results of redlining were devastating to minority and low-income areas. The lack of access to financial instruments like loans and insurance limited investment in the community, and in housing development and redevelopment. As a result, these areas experienced significant declines, especially in relation to the suburban areas which were benefitting from the financial investment preference.

Without access to home loans, minorities found it harder to purchase homes and build equity, and thus could not build wealth through home ownership. Many were forced to rent or enter into contract sales, which was a predatory lending approach combining the responsibilities of homeownership with the disadvantages of renting. Also, because loans were not available in these areas, the demand was lower and thus home prices were suppressed.

To understand the potential impacts of practices like redlining on insurance, consider homeowners insurance. One of the underwriting considerations in determining the insurability of a home is the maintenance and upkeep of the home. If, for example, the wiring in a home is knob-and-tube wiring and the roof is 20 years old, an insurance company is less likely to underwrite the home, and if coverage is offered, it may be at a higher cost and/or with lower coverage amounts. This is because homes with older, outdated wiring systems are more susceptible to fire losses, and homes with older roofs are more susceptible to roof damage or water losses during storms. In this case, the underwriting and rating decisions are based on the relationship of the property characteristics to risk of loss. But what if the reason that the electrical system and roof have not been updated is because the homeowner did not have the resources to pay for these updates, or financial institutions were less willing to lend money to the homeowner to make these updates? Ultimately, in this case, while the likelihood of loss is higher, the condition of the property is related, at least in part, to the history of disparate treatment.

While the practice of redlining is no longer allowed, the impact of over-a-century-old redlining practices is still being felt today in those communities which were discriminated against. You only have to look at the economic statistics of areas that were historically redlined to see the disparities in home values, income and wealth. Thus, the concern being raised is whether these historical practices are still influencing insurance practices and rating today. Said differently, even though redlining is not practiced today as it was in the 1930's, do insurance rating, underwriting and claims practices today produce similar outcomes, even though they are based on loss and other insurance outcome data? This would be the case if the underlying data being analyzed by insurance companies were being unduly influenced by the history of systemic racism.

The purpose of the methods discussed in this paper is two-fold First, we will examine potential approaches to identify the possible existence of discriminatory effects on protected classes in insurance rating, underwriting and claims processes today. Second, once we have determined the extent of the problem, we will discuss potential methodologies to mitigate the discriminatory effects.

2. Accusations of Bias in Insurance

While the calls to insurance companies for social justice have gotten louder as a result of 2020 events, the accusations of bias in insurance are not new. Listed below are examples of similar assertions over the past several years:

- In April 1997, the Center for Economic Justice released a report titled "Auto Insurance Redlining in Texas: Availability Worsens." This report alleged that private passenger automobile insurers in Texas were redlining by disproportionately rejecting drivers in poor and minority communities from their standard companies and placing them in sub-standard companies or the Texas Auto Insurance Plan.[1]

- In April 2017, ProPublica and Consumer Reports released a study claiming that minority neighborhoods paid higher auto insurance premiums than white areas

[1] http://www.cej-online.org/april97.php

with the same risk.[2] This study compared rate example premiums for 44 representative risks to loss experience obtained from departments of insurance. While there were methodological flaws in the analysis, as determined by a review of the data and study methodology by Pinnacle the article raised a number of questions about whether insurance rates were biased high in minority communities.

- In November 2017, the New York superintendent of financial services issued Insurance Regulation 150, which required insurers to demonstrate to the satisfaction of the superintendent that the use of education and occupation did not result in rates that are unfairly discriminatory.[3] As a result of the new regulation, insurance companies that were using education and occupation agreed to remove these two variables from consideration. Consumer advocates concluded from this agreement that insurance companies knew these factors were unfairly discriminatory, but ultimately the businesses were not confident that the use of education and occupation could be justified to the superintendent.

- In April 2019, the Algorithmic Accountability Act of 2019 was introduced in the United States House of Representatives. Though it ultimately did not pass, this act would have required companies using predictive models and machine learning algorithms to conduct an impact assessment of these models under rules to be established by the Federal Trade Commission. One required assessment to be completed would have been the determination of the risk of unfair, biased, or discriminatory decisions.[4]

Since 2020, there have been a number of efforts to address the potential of bias in insurance rates and processes. These efforts have been driven by insurance regulators, legislators, consumer groups and some insurance companies.

Insurance Regulators

Governed by the chief insurance regulators in 50 states, the District of Columbia and the five U.S. territories, the National Association of Insurance Commissioners (NAIC) provides expertise, data and analysis for insurance commissioners to effectively regulate the industry and protect consumers. The NAIC formed the Special Executive (EX) Committee on Race and Insurance in 2020. A workstream of one of the special (EX) committee' charges is to "examine and determine which practices or barriers exist in the insurance sector that potentially disadvantage people of color and/or historically

[2] ProPublica. "Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas with the Same Risk" https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk

[3] https://dfs.ny.gov/system/files/documents/2020/11/rf150a2txt.pdf

[4] https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202019%20Bill%20Text.pdf

underrepresented groups."[5] Hearings have been held to discuss the use of rating factors and whether some factors are unfairly discriminatory. The adopted charges for 2021 include "developing analytical and regulatory tools to assist state insurance regulators in defining, identifying, and addressing unfair discrimination in property/casualty (P/C) insurance."[6] The NAIC also has issued resolutions on the use of specific factors.

Regulators have also been involved at the state level. The most significant action by a regulator was the issuance of an emergency order in 2021 by Commissioner Mike Kreidler in the state of Washington banning the use of credit scores for three years.[7] This order was effective for all policies effective June 2021 or later, but was struck down by Thurston County Superior Court on October 8, 2021. Commissioner Kreidler is now pursuing a permanent rule to ban the use of credit based insurance scores.

Legislators

Legislators at both the federal and state level have also introduced legislation that would impact the rating of insurance.

In September 2020, United States Senator Cory Booker introduced Senate Bill 4755, the Prohibit Auto Insurance Discrimination (PAID) Act.[8] This act would have prohibited 12 factors from being used to price auto insurance, including credit-based insurance scores, gender, education and occupation. The purpose of this bill was described as "prohibiting insurance companies from using income proxies to determine insurance rates." This bill and the related House Bill 3693 were not enacted into law.

Legislation was introduced in multiple states that would have prohibited certain factors from being used in rating automobile and homeowners insurance (e.g., California is one of the states that already prohibited the use of insurance scores in rating). Four states where such legislation was recently introduced include:

- **Maryland**: prohibit the use of credit-based insurance scores.
- **Washington**: prohibit the use of credit-based insurance scores.
- **Oregon**: require automobile insurance companies to base rates only on driving record, years licensed, miles driven and other optional factors approved by the insurance commissioner.

[5] NAIC Special (EX) Committee on Race and Insurance. 2021/2022 Adopted Charges. https://content.naic.org/cmte_ex_race_and_insurance.htm

[6] NAIC Special Committee on Race and Insurance. 2020 Adopted Charges. https://content.naic.org/cmte_ex_race_and_insurance.htm

[7] https://www.insurance.wa.gov/sites/default/files/documents/r-2021-02-cr-103e_0.pdf

[8] https://www.congress.gov/bill/116th-congress/senate-bill/4755/text

- **Louisiana**: prohibit the use of credit-based insurance scores, education, occupation and gender from use in rating automobile insurance.

None of these bills passed, but one Colorado bill did pass that required insurance companies to demonstrate that their rating and insurance processes are not unfairly discriminatory. This requirement applies to most lines of insurance and all insurance processes, not just pricing. This bill will be effective on January 1, 2023, and the insurance commissioner will be working with the insurance industry and interested parties to determine how this requirement will be satisfied.

Consumer Groups

Organizations such as the Consumer Federation of America, the Center for Economic Justice, Consumer Reports and ProPublica have been calling for regulators and legislators to address bias in insurance. There are several consumer groups that have been calling for regulators and legislators to address bias in insurance. These groups include the Consumer Federation of America, the Center for Economic Justice, Consumer Reports and ProPublica. Examples of the calls from these organizations were provided earlier in this section. Consumer groups have continued to be active in this area.

Insurance Companies

A few insurance companies have been publicly calling for changes in the way that automobile insurance prices are determined. Examples of these calls include:

- Root Insurance announced in 2020 that they would be discontinuing the use of credit-based insurance scores by 2025. They have also called on other insurance companies to do the same.[9]

- Loop, a startup insurtech for auto insurance, has committed to not use factors such as credit-based insurance scores, education and occupation in rating insurance.[10]

- Sigo Seguros launched a Spanish-first automobile insurance product in Texas, and removed what the firm says are "biased rate factors like credit score, employment history and level of education."[11]

[9] "We're dropping credit score from car insurance pricing by 2025. Here's why." August 6, 2020. https://www.joinroot.com/blog/dropping-credit-score-from-car-insurance-by-2025/

[10] "Loop launches out of stealth to make auto insurance more equitable". January 13, 2021. https://techcrunch.com/2021/01/13/loop-launches-out-of-stealth-to-make-auto-insurance-more-equitable

[11] https://www.insurancejournal.com/news/southcentral/2021/08/04/625622.htm

Given the level of activity that has occurred since 2020, it is likely that this issue will remain an area of focus for legislators and regulators in the future.

3. What is Unfairly Discriminatory?

The key issue being debated as part of these discussions is whether rates and insurance processes in general, or certain rating factors specifically, are unfairly discriminatory or result in discriminatory effects on protected classes. However, the definition of these terms is still the subject of frequent debate. For a more complete discussion of these terms, please see "Defining Discrimination in Insurance," one of the pieces in this CAS Research Paper series authored by Kudakwashe F. Chibanda, FCAS. For purposes of this discussion, we will use the areas of focus described below.

From the perspective of actuarial standards, unfairly discriminatory generally refers to whether rates are supported by loss experience. Most states have laws in place that require that rates be "not inadequate, not excessive, and not unfairly discriminatory." The Statement of Principles Regarding Property and Casualty Insurance Ratemaking states 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 costs associated with an individual risk transfer."

Actuarial Standard of Practice No. 12 (ASOP 12), Risk Classification (for All Practice Areas), states that "rates within a risk classification system would be considered equitable if differences in rates reflect material differences in expected cost for risk characteristics. In the context of rates, the word fair is often used in place of the word equitable."

The Ratemaking Statement of Principles and ASOP 12 both focus on cost justification. However, there is a second consideration that is part of ratemaking today. ASOP 12 indicates that actuaries may have to deviate from the guidance of the standard to comply with applicable law. Many states have laws in place that prohibit insurance companies from using certain protected characteristics including race, religion and national origin in setting auto and homeowner insurance rates. In other words, there are factors that have been identified that, no matter how predictive they are of loss, cannot be used by insurers.

Even though there is still discussion on the exact definition of unfair discrimination and disparate impact, it is tied to the prohibition of the use of protected characteristics in setting rates. Consider the hypothetical example where an insurance company identifies a certain rating variable that is perfectly correlated with race. If race is predictive of loss, then the identified rating variable will also be predictive of loss. If the insurance company decided to use the variable in rating, it would technically not be in violation of the rating law, as the alternative variable is predictive of loss. However, it would not be compliant with the spirit of the law, as the predictive power of the alternative variable is perfectly correlated to a prohibited factor: race.

Ultimately, the question of discriminatory effects on protected classes (or unfair discrimination) comes down to, at least in part, whether individual factors or combinations of factors derive their predictive power in full or in part from their correlation with a prohibited characteristic. If so, then it must also be determined whether this results in disproportionately higher or lower rates for certain groups within that protected class.

While the level of correlation that is acceptable will not be decided by actuaries, they can be part of the discussion by determining the extent of the issue in insurance rating and by demonstrating the impact of various proposed solutions on rating and other insurance processes. Section 4 identifies approaches to identifying and adjusting rates to remove unfair bias.

4. Approaches for Measurement and Mitigation of Discriminatory Effects on Protected Classes

Given that many insurance processes are impacted by predictive models and machine learning, the approaches in this paper are described in the context of these methods. With minimal modification, these methods can be applied to the analysis of outcomes from most approaches. The essence of these approaches is to determine, after controlling for the distribution of policy characteristics, whether there is a discriminatory effect on protected classes.

Over the last few years, concern has been growing over the ever-expanding use of automated machine learning algorithms. Broadly stated, machine learning refers to the application of computational techniques for the purpose of analyzing historical data and using this information to make predictions about future events. These algorithms are designed to learn from data and predict future outcomes in various domains of social life including policing, sentencing and parole decision-making, medical diagnosis, facial recognition, loan approval, hiring practices, matchmaking, target advertising, and even risk management and pricing in insurance. Big data and sophisticated predictive algorithms have gained momentum and praise from the data science community as promising and efficient tools for providing solutions to myriad questions. However, it has also been discovered that such technologies, as well as the data they are trained on, can be fraught with inherent bias and discriminatory treatment, whether intentional or not, toward certain protected demographic groups in society.

The end result of utilizing biased data and employing predictive models built from such data, unfortunately, can propagate that bias into the choices made from the models' predictions. For this reason, researchers from the field of artificial Intelligence (AI) and machine learning have embarked on a mission to design and improve methodologies for detecting and mitigating bias, aiming to preserve the intended salutary impact of the models' outcomes as well as construct systems that are potentially more equitable.

Fully or semi-automated systems for making predictions about patterns in people's lives (referred here simply as "models" or "classifiers" in the binary case) appear to be designed and applied fairly and consistently. They are, however, inherently capable of introducing unfairness into the process and thus have direct consequences to individuals affected by these models. Bias is all around us, and it can creep into the decision-making paradigm in subtle ways, whether it is the subjectivity of human judgement, prejudice, historical inequities baked into the data, or faulty algorithms. As a result, models can sometimes result in unreliable and unfair decisions by uncovering correlations in the data that are only reflections of those biases and historical inequities.

There also exists a mathematical kind of bias, which in statistical parlance refers to the degree to which an estimator, on average, deviates from the true value of the parameter it is intended to predict. For example, if the true value of the claim severity of a subset of an insurer's portfolio is $1,000, but the model estimates severity to be $1,500, then the statistical bias for that specific model is $500, overshooting the ground truth by +50%. Minimizing bias is therefore a desirable characteristic of models, as it substantiates that the model has correctly captured the relationships among the variables in our data. It is important to note that statistical bias is analyzed independently from the legal and ethical concept of fairness, and is simply a mathematical property of the modeling algorithm.

But while a model can be unbiased in the statistical sense, given all precautions have been taken to ensure accurate model specification and relevant feature selection, it may still not be enough to apply such a model by itself in real-life scenarios where human interests are at stake. Even if that statistical bias is minimized, demographic disparities existing in society as well as bias in human cognition that may be woven into historical data could still make their way through the modeling process and affect the model's output.

The latest research in model fairness and model de-biasing suggests introducing an additional component to the concept of model bias that transcends the purely statistical context. The central theme in this additional dimension of bias detection and bias mitigation attempts to provide practitioners of analytics with mechanisms and mathematical constructs to minimize the social inequalities that their models may capture through data and ensure that the model does not unfairly discriminate against certain protected classes. The challenging aspect of managing both statistical bias and demographic bias is that minimizing one will often increase the other; any intervention to the data or the model should consider the effects from this interaction. To balance these competing priorities, we must first understand the extent to which any disparities created by the actions taken based on the model are unfair and socially unacceptable, as well as how bias-mitigating efforts to ameliorate the model affect the model's accuracy across the protected classes.

Predictive models can be utilized for two different forms of estimation — one is regression, and the other is classification. In the regression context, the model estimates a numeric response, such as a customer's insurance premium, while in classification, it predicts one of several discrete categories, with the binary case being the more common scenario in the insurance field. Examples of binary classifiers with insurance applications include predicting whether a claim is fraudulent, predicting a customer's risk of leaving, or simply predicting the occurrence of a claim. The types of fairness metrics discussed in the next section are equally applicable to both regression and classification models, but for simplicity of exposition, we will limit ourselves to the binary case.

We will use Y to denote the response variable, with 1 indicating a positive outcome, and 0 indicating a negative outcome. The positive outcome usually denotes the class that we are interested in predicting, e.g., the occurrence of a claim would be labeled as 1 and treated as the positive class. In many applications, the positive class comprises a much smaller proportion of the training data — in personal auto insurance, for example, only a small percentage of insureds file a claim. This type of imbalance between the two output categories poses its own challenge to the model's ability to produce equally accurate predictions for the various groups of insureds. We will also assume there is a single protected or sensitive attribute, labeled as A, with two different subclasses, a and b. In the examples discussed in this section, the protected attribute referenced is gender.

Like all predictive algorithms, binary classification models are not perfect and are prone to prediction error. There are two types of statistical mistakes a model can commit: Type 1 and Type 2. Type 1 error, also known as a false positive rate, occurs when the model predicts a claim — the positive outcome — for a group of policies that did not actually experience a claim. Similarly, a false negative rate, or Type 2 error, indicates an error in the opposite direction, failing to predict a claim for a group of policies that did experience a claim. In combination, the size of these two errors determines the overall accuracy of the classifier as the overall fraction of correctly predicted outcomes. The mathematical definitions of the fairness criteria that have been proposed incorporate requirements of varying degree of strictness for these errors by enforcing additional restrictions on the joint distribution of the sensitive attribute, the response variable, and the model's predictions.

Broadly speaking, the taxonomy of fairness includes three fairness measures classifications: independence, separation, and sufficiency.

Figure 2: Categories of Fairness Criteria

Independence	Separation	Sufficiency

A - protected attribute

Y - observed value of target variable

Ŷ - predicted value of target variable

Independence represents the simplest and most intuitive categorization that focuses only on the distribution of the model's predictions across the various demographic classes and requires that the predictions and the protected attribute be statistically independent. Separation goes one step further and also considers the observed values of the response variable. The criterion of separation is satisfied if the predictions and the protected attribute are statistically independent but conditional on the actual response. With separation, the model's predictions are allowed to vary across the attribute classes as long as the actual response values are different. Sufficiency is similar to separation, except that the comparisons are conditional on the same predicted values instead of the same observed responses.

Fairness Definitions in Modeling

The following list of fairness definitions is not exhaustive but rather attempts to summarize some of the most popular criteria promoted in the research literature. While outside the scope of this paper, it is important to mention that many of the fairness metrics are mutually incompatible and hence cannot be satisfied simultaneously.

Definitions Based on the Model's Predictions (Independence)

- **Demographic Parity**. Demographic parity, also known as statistical parity, or equality of outcomes, simply requires that the model makes equal predictions for both classes. For example, if the purpose of the model is to predict the occurrence of a claim and the protected attribute is gender, then this criterion will be satisfied if the model produces the same probability of claim incidence for both groups. In mathematical notation, this can be expressed as follows:

  $P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b)$.

  While this criterion enforces independence between the algorithm and the protected attribute, its main disadvantage is that it completely ignores the algorithm's accuracy rates for each protected class. It is easy to see that we can build a model that accurately picks up the correlation between, say, males and their rate of claim occurrence, but also makes a random assignment within the group of females, making sure to achieve the same rate of claim occurrence, thus still satisfying this fairness metric.

  A slight variation of demographic parity is conditional demographic parity, which allows for other non-sensitive attributes to affect the predicted outcome. Such attributes could include years of driving experience and vehicle age. This definition is satisfied if the protected and unprotected classes are assigned the same predictions after controlling for the permitted factors. In the claim occurrence example, the requirement is met if males and females have the same predicted probabilities given both groups have the same driving experience and drive vehicles of identical age.

One of the early proposed methods to solve for independence is the so-called ″fairness through unawareness,″ which simply reflects the situation where sensitive features are not explicitly used in the modeling algorithm; hence, any decisions based on the model were thought to be independent from those features. If a predictive model for frequency entirely excludes the gender variable, then that model is said to satisfy fairness through unawareness. Unfortunately, simply removing the sensitive attribute from the model is effective only when the sensitive attribute is independent from any other variables included in the model, which occurs rarely in practice. Often, as a result of existing correlations among the variables, other variables in the model can serve as proxies for the sensitive attribute and thus indirectly impair the independence criterion.

Definitions Based on the Model's Predictions Conditional on Actual Outcomes (Separation)

- **Equal Opportunity**. Equal opportunity is an extension of demographic parity, which still requires that the predicted outcomes are equal across the protected classes, but is conditional on the positive outcome being observed. This criterion represents an improvement over the simple demographic parity because the introduction of this second condition ensures that now the true positive rates are equal for both groups. In our example, males who experienced a claim will be assigned the same probability of claim as females who also experienced a claim, and consequently, the model will exhibit the same misclassification rate for the positive outcome of both groups. While equal opportunity allows control of the false negative rate, it does not guarantee equality of the false positive rates across the classes. Mathematically, we have:

  $P(\hat{Y} = 1 \mid Y = 1 \;\&\; A = a) = P(\hat{Y} = 1 \mid Y = 1 \;\&\; A = b)$.

- **Equalized Odds**. The equalized odds[12] measure improves upon equal opportunity by imposing the stricter requirement that the attributes' classes have equal true positive rates and equal false positive rates. Equivalently, the false negative rates and true negative rates must be the same across the two groups. In our example, this implies the following:
  - the probability that a policyholder who actually experienced a claim was correctly predicted to have a claim.
  - the probability that a policyholder who did not experience a claim was incorrectly predicted to have a claim.
  - the above probabilities should be the same for males and females. Using the above notation:

[12] Also known as Disparate Treatment Avoidance.

$P(\hat{Y} = 1 \mid Y = y \;\&\; A = a) = P(\hat{Y} = 1 \mid Y = y \;\&\; A = b),\ y \in \{0, 1\}$.

Because this criterion is stricter in nature, it generally results in a model having lower overall accuracy. Also, it is evident that if an algorithm satisfies equalized odds, and the two attribute classes have inherently different rates of claim occurrence, the algorithm will necessarily produce different precision rates for each class, i.e., the proportion of correct positive predictions will be different for the different gender classes.

Other types of separation metrics similar to equalized odds include conditional procedure accuracy, overall accuracy equality and treatment equality. The overarching idea with respect to all these metrics is that unlike the independence criteria, they allow for a non-zero correlation between the protected feature and the model's predictions to the extent that the response variable indicates such differences. Unlike demographic parity, we are not insisting that males and females have the same predicted probability of claim, but rather they experience the same model error rates.

Definitions Based on the Actual Outcomes Conditional on Model's Predictions (Sufficiency)

Calibration. Calibration[13] is the most popular sufficiency metric which requires that, conditional on the same predicted probability score *p* by the model, both the protected and unprotected classes have the same probability of actually belonging to the positive outcome. This criterion is very similar to requiring equal precision rates for both classes except that with precision, the decision is made after the application of a preselected threshold, while calibration is more general, essentially requiring the same precision rate for every possible threshold. Mathematically, we have:

$P(\mathrm{Y} = 1 \mid P = p \;\&\; A = a) = P(Y = 1 \mid P = p \;\&\; A = b),\ p \in [0, 1]$.

In the claims occurrence example, calibration means that whenever males and females have the same predicted probability of incurring a claim, their respective, actual observed claim rates are also the same. For instance, if a group of males and a group of females both have an estimated claim probability of 80%, and both show an actual claim frequency rate of around 70%, then the model satisfies calibration. But if the actual claim frequency of males is 70% and that of females is only 45%, the model would be deemed unfair against females when $p = 80\%$ because the model would disproportionally misclassify females as high risk.

[13] This also goes by the name of Test Fairness.

Well-calibration. Well-calibration extends the definition of calibration by including the additional requirement that for a given predicted probability score *p*, the actual observed proportions should also equal *p*. If a model determines that males and females have the same probability of claim occurrence, say 80%, then their actual frequency rates should also equal 80%, and this relationship should hold for all possible values of the probability score. Mathematically, we have:

$P(Y = 1 \mid P = p \;\&\; A = a) = P(Y = 1 \mid P = p \;\&\; A = b) = p,\ p \in [0, 1].$

To put the above fairness definitions in context, let's assume we have built a frequency model with the following track record:

- The model is more likely to predict a claim-free status for males than females, so it does not satisfy demographic parity.
- The model is more likely to predict a claim-free status for males who have actually experienced a claim than females with similar experience, so it does not satisfy equalized odds.
- The model provides an overall equal prediction accuracy for both genders, i.e., regardless of gender, the model is equally accurate for insureds who have experienced a claim and those who have not.
- The model applies the same treatment according to the equal opportunity metric to males and females who have not experienced a claim, assigning the same true positive rate and the same false negative rate for both.

Would such a model be considered fair given the above results? The answer to this question clearly depends on which metric or combination of metrics we believe are most appropriate and will most certainly require the concerted effort and continued dialogue among policymakers, actuaries, regulators and consumer advocates to reach a reasonable consensus. In their paper, "Algorithmic Fairness: Contemporary Ideas in the Insurance Context," Dolman and Semenovich state: "It would be prudent to act in four related ways:

1. Create Internal Clarity — we should be clear on why we consider our actions to be fair and reasonable;

2. Acknowledge Imperfections — we should acknowledge the inherent tradeoffs required;

3. Be Adaptable — we will likely need to adapt any answer over time, particularly as the research environment matures; and

4. Act With Humility — commitment to openly discuss views which may contradict our own, and a commitment to rectify any issues as they are identified, and adapt

> according to society's evolving norms, appears the only reasonable course of action."

Bias Mitigation Techniques

Once we have identified potential bias, the next step is to determine how to mitigate it. The answer to this question will not be answered solely by actuaries, but one of the answers may be to adjust model outcomes (i.e., indicated rating factors) to remove an identified bias. This section introduces these mitigation techniques. We refer the reader to the references for a more complete treatment of these mitigation methods.

Given a set of fairness criteria, a variety of bias mitigation techniques can then be applied to satisfy these criteria. The initial goal in the bias mitigation process is to determine whether the training data discriminates against a given protected attribute with respect to the selected target on the basis of the pre-selected bias metrics. If these metrics score above or below an acceptable threshold, we proceed with applying one or more mitigation algorithms to restore equity across the groups. The bias mitigation algorithms attempt to improve the fairness metrics by modifying either the training data, the model or the predictions themselves, depending on where exactly in the modeling lifecycle the bias has been detected. The various de-biasing approaches can be broadly classified into three principal categories: fair pre-processing, fair in-processing, and fair post-processing, with each encompassing a library of individual algorithms.

The first class of de-biasing techniques belongs to the so-called fair pre-processing methods, where the goal is to remove any underlying bias from the data prior to modeling. Algorithms in this class, such as re-weighting and disparate impact remover, ensure that the input data is fair and balanced, so that the predictor space is uncorrelated with the sensitive attribute. This can be achieved by changing the class labels of the data set, and by re-weighting or re-sampling the data.

The fair in-processing group of techniques performs various modifications to the learning algorithm in order to minimize the discrimination during the model training, either by incorporating changes into the objective loss function or imposing the fairness constraint directly into the optimization process. The prejudice remover, for example, introduces a fairness penalty term directly to the loss function, similar to traditional regularization modeling techniques. The regularization parameter used in this method controls the trade-off between predictive accuracy and the degree of fairness. Generative adversarial networks are some of the most recently developed analytical tools for fair classification using adversarial de-biasing. In this context, a neural network classifier is trained in the traditional way, but at the same time, the ability of the adversarial neural network to predict the protected attribute is minimized, which then ultimately minimizes the correlation between the sensitive information and all other information.

The final category is fair post-processing methods, which intervene in the final stage of the analytical pipeline after the model has been trained. These methods operate directly on the model predictions, so that they are uncorrelated with the protected attribute. One approach is the Bayes optimal equalized odds predictor, which changes the predicted labels based on the equalized odds fairness metric. Other examples include reject option classification and calibrated equalized odds. The most important advantage of these algorithms is that fairness can be achieved without having to re-train the original model.

Example

To contextualize the fairness metrics and de-biasing techniques discussed above, the following example illustrates their practical application to an insurance dataset and a simple, pricing Generalized Linear Model (GLM) we have built using the data. The modeling dataset we used is part of the ***fremotorclaim*** database, which is available in the public domain (https://fairmodels.drwhy.ai/) and can be downloaded from the CASdatasets package in the open-source environment R. This dataset contains anonymized information from a French private passenger automobile portfolio with observed policy-related characteristics and respective loss experience for the years 2003 and 2004.

The focus of this basic example is only to illustrate the concepts; it is not intended to make recommendations on any specific course of action or analysis regarding data manipulations, model building, or bias detection and mitigation steps.

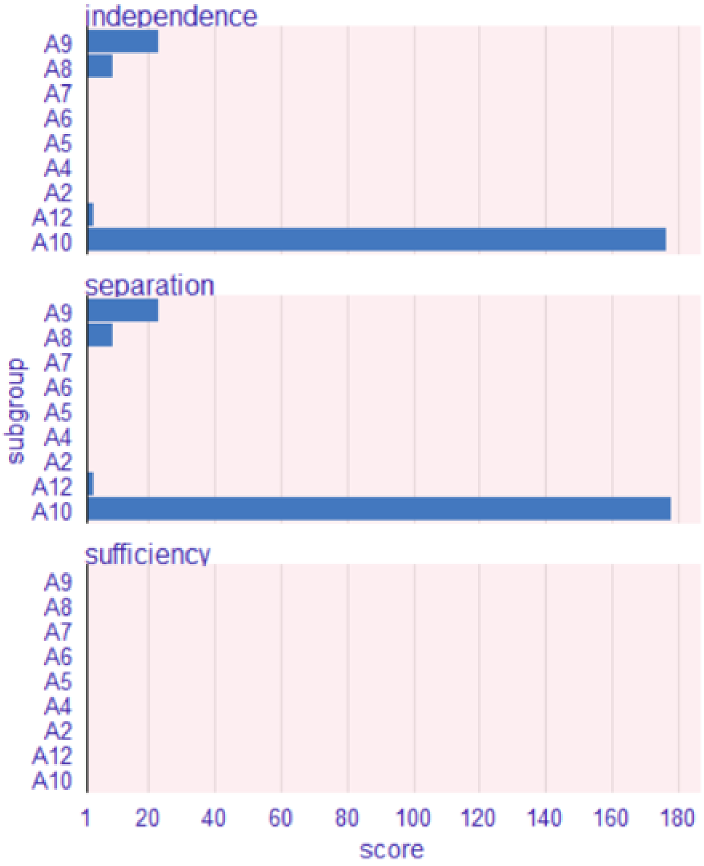
We begin by building a simple GLM, aiming to predict the frequency of claims, conditional on the values of the policy characteristics that we use as predictors in our model. We apply the traditional assumptions that the number of claims follows a Poisson distribution, and that the individual observations are independent. We have selected area as the protected attribute of interest, a nominal variable with 10 levels (A2-A10, A12), representing the geographic region of the policy. For this example, we will assign the privileged status to area A3, evaluate the fairness criteria for bias with respect to our GLM, and try to de-bias the GLM by applying the re-weighting technique from the in-processing group of techniques.

To perform the bias detection and mitigation, we resorted to the utilities in the ***fairmodels*** package, also available in R. This package is appropriate to our task at hand as it comes equipped with fairness metrics that can be applied not only to binary classification problems but also to the more general regression case, where the target to be predicted is either multi-level or continuous. The fairness measures discussed above have been studied and evaluated most extensively for binary targets but have been recently extended to the regression case and included in the ***fairmodels*** package.

In the first step, we built a model that uses all available predictors, including area. Next, we test the fairness criteria based on the model's output using the 80% rule, which requires that the value of the metric for the unprivileged group be at least 80% of that for the privileged group. In other words, the ratio of the privileged metric value to the unprivileged metric value should ideally be 1.00 — evidence the model has not learned from the protected attribute — and no greater than 1.25.

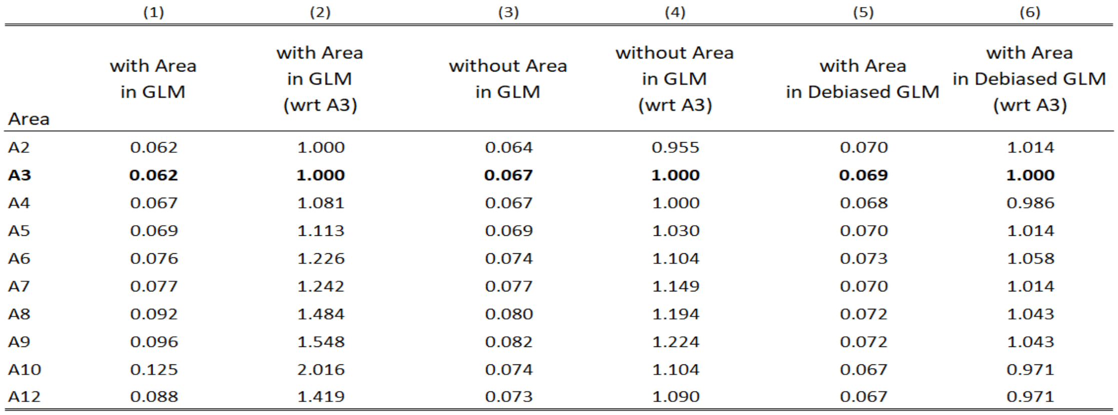
As can be seen in Figure 1, of the three fairness metrics — independence, separation, and sufficiency - only sufficiency is satisfied. According to independence and separation, bias is detected for most areas, but a rather outsized bias is detected for areas A8, A9, and A12, and especially for area A10. For these areas, the metrics far exceed the threshold of 1.25. Based on these results, we can conclude that the model's output is not independent from the protected attribute.

Figure *3*: Bias Detection Metrics for GLM — With Area in GLM



To further investigate the bias, we calculate the average predictions by the model for each area (Table 2). In columns (1) and (2), we show the raw average predictions and the respective average relativities, re-based with respect to area A3. The areas with detected bias show average relativities much greater than 1.00 with respect to area A3, and policies from these areas are therefore charged a much higher premium on average.

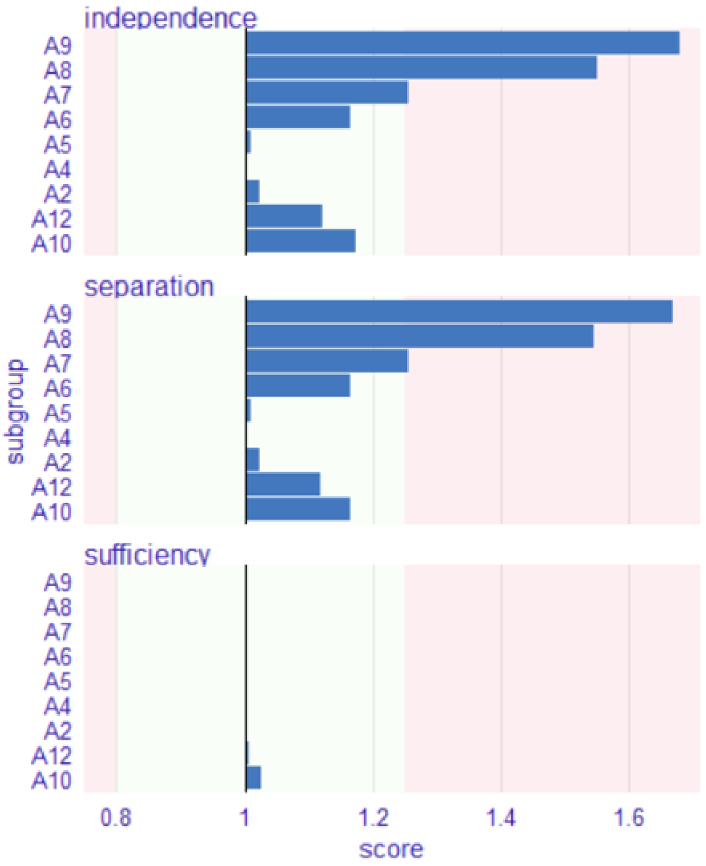
Figure 4: GLM Average Predictions by Area



	(1)	(2)	(3)	(4)	(5)	(6)
Area	with Area in GLM	with Area in GLM (wrt A3)	without Area in GLM	without Area in GLM (wrt A3)	with Area in Debiased GLM	with Area in Debiased GLM (wrt A3)
A2	0.062	1.000	0.064	0.955	0.070	1.014
A3	**0.062**	**1.000**	**0.067**	**1.000**	**0.069**	**1.000**
A4	0.067	1.081	0.067	1.000	0.068	0.986
A5	0.069	1.113	0.069	1.030	0.070	1.014
A6	0.076	1.226	0.074	1.104	0.073	1.058
A7	0.077	1.242	0.077	1.149	0.070	1.014
A8	0.092	1.484	0.080	1.194	0.072	1.043
A9	0.096	1.548	0.082	1.224	0.072	1.043
A10	0.125	2.016	0.074	1.104	0.067	0.971
A12	0.088	1.419	0.073	1.090	0.067	0.971

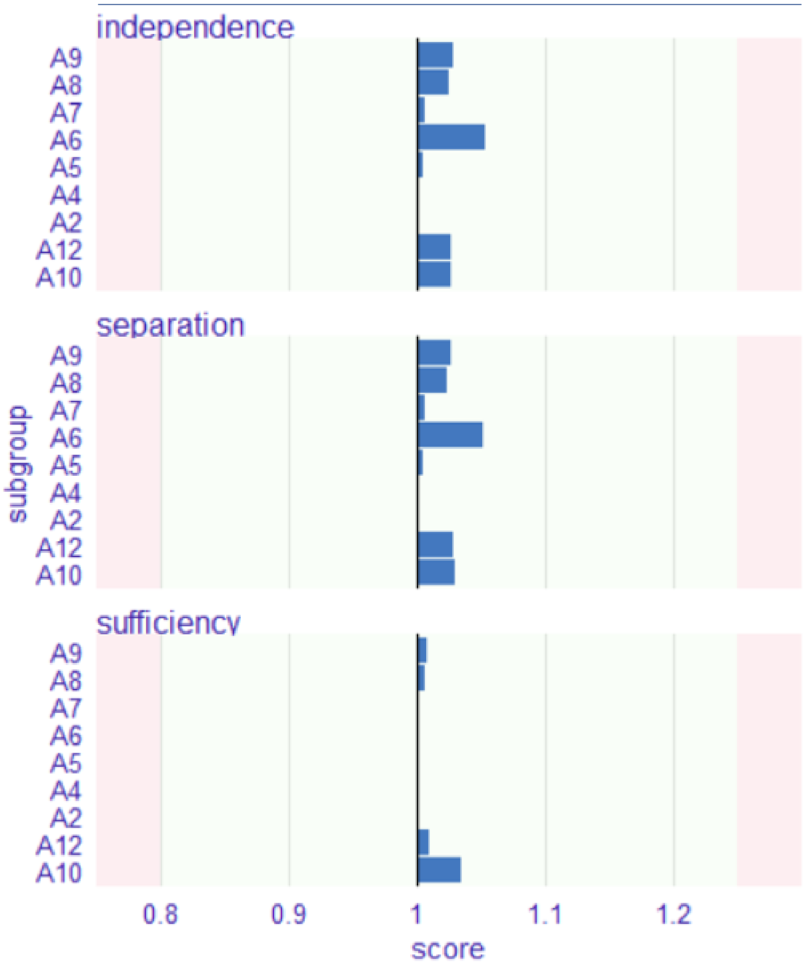
Next, in an attempt to remove the bias from the model's predictions, we apply the fairness through unawareness principle, and simply remove area from the model. Figure 2 shows that bias has been greatly reduced but not completely removed for a subset of the areas. In addition, column (4) in Table 2 confirms these results, with average relativities for areas A7, A8, and A9 still much higher than 1.00. These results indicate that simply removing a protected attribute from the model does not necessarily produce unbiased predictions with respect to the protected attribute.

Figure 5: Bias Detection Metrics for GLM — Without Area in GLM



Lastly, we test one of the more intuitive in-processing techniques — re-weighting, which assigns each data sample a weight so as to de-correlate the protected information from the permissible information. While it is not guaranteed that a single model de-biasing method will always produce fair algorithms, in our simple example re-weighting proved successful. Figure 3 demonstrates that now all three metrics are much less than 1.25. Additionally, column (6) in Table 2 shows that all average relativities are close to 1.00 because the model's predictions were forced to ignore area as a protected variable. All areas now are charged approximately the same premium, hence our pricing GLM has been de-biased in relation to area.

Figure 6: Bias Detection Metrics for GLM — With Area in De-biased GLM



Next Steps

The calls for eliminating bias in rating described earlier have generally not involved statistical measures of the alleged bias but have typically moved to proposed solutions (most generally the limitation or removal of certain rating variables). Unfortunately, without any measures showing the extent of existing bias, it is impossible to determine if these proposed solutions actually address the perceived problem. The methods outlined in this research brief provide a framework for quantifying discriminatory effects on protected classes. Once we understand the extent of the potential problem, appropriate solutions can be developed.

While the ability to directly test predictive models for bias as well as apply adjustments to the training data or model predictions to remove any detected bias represents a significant stride toward achieving algorithmic fairness, it is equally important that we strive to build models that are also transparent and easily interpretable. The easier it is to communicate and explain a model's decision-making process to the various stakeholders, the easier it will be to engender their trust in our algorithms. The domain of human interpretable machine learning, also known as “explainable AI,” is one of the most recent advances in the field of machine learning, in which the main objective is to help better understand: (a) how the model makes decisions, (b) which features drive the model’s predictions, and (c) how each explanatory variable contributes to the individual predictions. Bias detection and mitigation, in concert with enhanced model “explainability,” both from a global and local perspective, are likely to represent the next frontier of building more efficient and more equitable algorithms. The actuarial community is well positioned to rise to the challenge and play a role in promoting fairness and social justice.

*****

Research and education are vital to the success and evolution of the Casualty Actuarial Society (CAS), the actuarial profession, and the broader insurance industry. As the industry discourse on potential bias in insurance pricing evolves, the CAS will continue to develop resources to support members and industry professionals and is open to collaborating with others. As the CAS pursues further research and educational opportunities and the development of new approaches to address these issues, we invite anyone interested in collaborating with the CAS on future research or educational sessions to reach out by sending an email to diversity@casact.org.

References

[1.] Barocas, S., Hardt, M., Narayanan, A. (2021). Fairness and Machine Learning: Limitations and Opportunities.

[2.] Corbett-Davies, S., Goel, S. (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning,* Stanford University, Association for the Advancement of Artificial Intelligence.

[3.] Dolman, C., Semenovich, D. (2018). Algorithmic Fairness: Contemporary Ideas in the Insurance Context, actuaries.org.uk.

[4.] Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., Venkatasubramanian, S. (2015). *Certifying and removing disparate impact,* Haverford College, University of Utah, and University of Arizona.

[5.] Fletcher, R., Nakeshimana, A., Olubeko, O. (2021). Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health, Frontiers in Artificial Intelligence.

[6.] fremotorclaim database, http://dutangc.perso.math.cnrs.fr/RRepository/.

[7.] Ghassami, A., Khodadadian, S., Kiyavash, N. (2018). *Fairness in Supervised Learning: An Information Theoretic Approach,* University of Illinois at Urbana-Champaign, Association for the Advancement of Artificial Intelligence.

[8.] Kamiran, F., Calders, T. (2015). Data preprocessing techniques for classification without discrimination, Springerlink.com.

[9.] Mahoney, T., Varshney, K., Hind, M. (2020). AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning, O'Reilly.

[10.] Wiśniewski, J., Biecek, P. (2021). *fairmodels: A Flexible Tool For Bias Detection, Visualization, And Mitigation*, arxiv. https://arxiv.org/abs/2104.00507.

[11.] Zhang, B., Lemoine, B., Mitchell, M. (2018). *Mitigating Unwanted Biases with Adversarial Learning,* Stanford University, Google, and Association for the Advancement of Artificial Intelligence.