Award Winner

Beyond Redlining: Addressing Potential Zip Code Bias in Life Insurance Pricing

Joshua Owusu

October 2025

The views and ideas expressed in this essay are the author's alone and do not represent the views or ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, or the author's employer.

Introduction

Insurance models, whether for pricing, risk assessment, or customer engagement, commonly rely on variables such as age, gender, and location. While these factors can be useful in making predictions about outcomes like mortality and morbidity, they may not be free from ethical concerns. In particular, geographic rating variables like ZIP codes, though appearing neutral, can reflect socioeconomic and racial disparities due to historical redlining.[1]

Although redlining was outlawed in 1968 by the Fair Housing Act, its legacy could continue to influence the datasets used in insurance, potentially introducing biases that impact decision-making.[2]

This essay discusses how ZIP codes may introduce bias in insurance datasets, focusing on their indirect influence on data used for life and annuity products. It then explores adversarial debiasing as a promising machine learning approach to mitigate these effects.

Understanding ZIP Code Bias

The history of ZIP code bias dates back to the 1930s, when the Home Owners' Loan Corporation created maps that labeled Black-majority neighborhoods as "hazardous" for investment.[2] Even when race may not be directly used, today's models may still learn biased patterns from historical data. By linking certain locations to higher risk, these models can unintentionally repeat past discrimination.

For life and annuity insurance products, the connection between ZIP code and bias in datasets, while often indirect, is still significant. Studies have found that redlined areas have comparatively fewer healthcare facilities and higher

[1] Redlining refers to systematic denial of services (for example, loans or mortgages) based on location without considering the qualifications of the individual applicant.

[2] Aaronson, Hartley, and Mazumder, (2021 November), "The Effects of the 1930s HOLC 'Redlining' Maps," American Economic Journal: Economic Policy 13 (4): 355–92. https://doi.org/10.1257/pol.20190414.

Caveat and Disclaimer

The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries Research Institute, the Society of Actuaries or its members. The Society of Actuaries Research Institute makes no representation or warranty to the accuracy of the information.

Copyright © 2025 Society of Actuaries Research Institute

mortality rates, following decades of disinvestment.[3, 4] Consequently, an algorithm trained on datasets from such areas could attribute higher risk to all individuals from those locations. Moreover, reliance on ZIP codes or their proxies can lead to missed opportunities for insurers by inaccurately tagging valuable clients from certain locations as high risk.

However, just like the age variable, (where a healthy 40-year-old might be lower risk than an unhealthy 30-year-old), ZIP codes can be misleading indicators of personal risk, as they may reflect systemic disadvantages more than individual health or behavior.

Adversarial Debiasing: An Algorithmic Ethics Coach

Adversarial debiasing offers an approach to mitigating bias within datasets. The process begins by training a primary model to predict a target outcome (Y) using inputs (X_i) while simultaneously training an adversarial model to predict a sensitive variable (Z) like ZIP code from the primary model's output.[5] Through multiple training cycles, the system penalizes the primary model whenever the adversary successfully predicts Z, gradually forcing it to develop fair representations.

The optimization process balances the two competing objectives through a total loss function $L = L_y - \alpha L_d$, where L_y represents the primary loss and L_d represents the adversary's loss[6] (α is a hyper parameter that controls the trade-off between the two objectives). A higher α prioritizes fairness and tries to prevent the adversary from predicting the sensitive variable correctly. A lower α favors prediction accuracy even if some bias remains.

Optimal hyper parameters (such as the alpha value in this case) can be selected by testing a range of values using cross-validation.[7] For each alpha, the model is trained and evaluated using performance metrics (like accuracy or Root Mean Squared Error). The final alpha is carefully selected to ensure that the model has the right balance between the insurer's need for accuracy and fairness across different demographic groups.

This technique has been applied to reduce inequality in several non-insurance contexts. For example, it helped address bias in COMPAS, a tool used in U.S. courts to predict reoffending, which often labeled Black defendants as high risk.[6] In another study using COVID-19 diagnosis predictions, adversarial debiasing helped reduce unfair differences in hospital data across different locations.[5] These cases demonstrate its potential to prevent variables like ZIP codes from acting as proxies for race in insurance datasets.

Other debiasing methods, such as reweighting and preprocessing, aim to tackle bias by adjusting the dataset prior to training. However, they can sometimes fall short of reducing the influence of proxy variables in the data.[8]

By applying adversarial debiasing, insurers can create models that maintain actuarial integrity while reducing unfair geographic biases.

[3] Lynch et al., (2021 June) "The Legacy of Structural Racism: Associations between Historic Redlining, Current Mortgage Lending, and Health," *SSM – Population Health*, Vol. 14, https://doi.org/10.1016/j.ssmph.2021.100793.

[4] Krieger et al.,(2020 July) "Structural Racism, Historical Redlining, and Risk of Preterm Birth in New York City, 2013-2017," *American Journal of Public Health* 110, 1046–1053, https://doi.org/10.2105/AJPH.2020.305656.

[5] Yang et al., (2023) "An Adversarial Training Framework for Mitigating Algorithmic Biases in Clinical Machine Learning," *npj Digital Medicine* 6, 55, https://doi.org/10.1038/s41746-023-00805-y.

[6] Wadsworth, Vera, and Piech, (2018) "Achieving Fairness through Adversarial Learning: An Application to Recidivism Prediction," arXiv:1807.00199, https://doi.org/10.48550/arXiv.1807.00199.

[7] James et al., (2013), "An Introduction to Statistical Learning: with Applications in R," Springer, DOI 10.1007/978-1-4614-7138-7, https://www.statlearning.com/.

[8] Wongvorachan et al., (2024) "A Comparison of Bias Mitigation Techniques for Educational Classification Tasks Using Supervised Machine Learning," *MDPI*, 10.3390/info15060326.

The Case for Adversarial Debiasing

Regulatory Compliance

Concerns over potential discrimination have triggered discussions about stricter oversight in the United States, similar to measures adopted in Europe.[9] While it does not explicitly mention ZIP codes, the new EU AI Act emphasizes the need for unbiased data in AI applications. Article 10 requires that datasets used for training, validation, and testing be representative, and examined for biases that could lead to discrimination.[10]

Adversarial debiasing offers a proactive solution, helping ensure that insurance datasets and the models built upon them align with emerging fairness regulations.

Reputation Management

Insurers can demonstrate their commitment to equitable practices, increasing consumer trust and improving brand image. In an era of public awareness around algorithmic bias, consumers are increasingly drawn to companies that prioritize ethical decision-making[11]. By adopting adversarial debiasing, insurers show a commitment to fair data practices. This sets them apart from competitors that are slower to make such changes.

Model Transparency

The process of adversarial training can make it easier to audit and interpret the role of various variables, including proxies like ZIP codes. By explicitly identifying and minimizing the model's reliance on them during training, adversarial debiasing can help provide a structured approach for uncovering hidden sources of bias. This clarity supports regulatory compliance efforts and facilitates internal model validation, enabling actuaries and data scientists to better justify geographic differences in datasets.

The Limits of Adversarial Debiasing

Adversarial debiasing has some drawbacks. First, it requires more computing power and training time because the system must balance two competing goals: accuracy and fairness. Additionally, designing an effective adversary and selecting hyper parameters require careful consideration, as overly aggressive models can remove useful signals in geographic trends. This can reduce the overall performance and accuracy of the model.[12] Third, the method works best with large, balanced datasets, which smaller insurers may not have.

Despite these obstacles, adversarial debiasing remains one of the best ways to increase fairness in insurance datasets without compromising actuarial rigor.

[9] Frees and Huang, (2021) "The Discriminating (Pricing) Actuary,"SSRN, http://dx.doi.org/10.2139/ssrn.3592475

[10] European Parliament & Council of the European Union, (2024 June 13), *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (the "Artificial Intelligence Act")*. *Official Journal of the European Union, L 1689*, 12 July 2024, https://data.europa.eu/eli/reg/2024/1689/oj.

[11] Sepideh Ebrahimi et al., "Reducing the Incidence of Biased Algorithmic Decisions through Feature Importance Transparency: An Empirical Study," *European Journal of Information Systems* 34, no. 4 (2025): 636–64, https://doi.org/10.1080/0960085X.2024.2395531.

[12] Members of the CAS Race and Insurance Pricing Research Task Force (2025) *Practical Application of Bias Measurement and Mitigation Techniques in Insurance Pricing: Part 2 - Advanced Fairness Tests, Bias Mitigation, and Non-Modeling Considerations*, Casualty Actuarial Society, https://www.casact.org/sites/default/files/2025-01/Practical_Application_of_Bias_Measurement_Part_2.pdf.

Copyright © 2025 Society of Actuaries Research Institute

Conclusion

Adversarial debiasing offers a forward-looking solution to address potential unfairness within insurance datasets. Since using ZIP codes can reflect past discrimination, insurers need better tools that are both accurate and fair. This method helps reduce potential ZIP code bias by teaching models to learn without relying on sensitive geographic information.

* * * * *

Give us your feedback!
Take a short survey on this report.
Click Here
SOA Research INSTITUTE

Copyright © 2025 Society of Actuaries Research Institute