Code Overview

Our implementation evaluated and mitigated unfair bias in our insurance claim prediction model using Microsoft's FairLearn package. After excluding sensitive features like gender and income, we measured fairness using Demographic Parity and Equalized Odds metrics, finding minimal but present disparities (0.0017 and 0.0046 respectively) with an initial accuracy of 0.7467. Our mitigation techniques successfully reduced these disparities while maintaining reasonable performance, with only a slight decrease in accuracy to 0.7443, effectively balancing fairness with model utility.

Overview of Fairness Metrics & Mitigation

Predictive models are considered best when their predictions are accurate. But what if this accuracy comes at the cost of bias? Therefore, it is essential to not only assess the overall performance but to also ensure that the results are not disparate towards certain groups. This becomes more critical when we deal with sensitive, protected features like gender and age - features that have been historically subjected to bias.

Bias in predictive models can be evaluated via fairness metrics that quantify how equitably a model performs. It helps answer questions like “Are error rates, benefits and risks distributed fairly?” By pinpointing specific inequities, these metrics guide developers in selecting and tailoring interventions like adjusting threshold values and reweighting.

The recall parity metric, also called Equality of Opportunity, focuses on achieving the same true positive rate across groups, ensuring that individuals who deserve a positive outcome are equally likely to be identified by the model irrespective of their background. The false discovery rate metric ensures that the rate of false positives - instances when the model incorrectly predicts a positive outcome - is similar across groups. The equal odds metric certifies that the error rates, both false positives and false negatives, should be similar for all groups, thus balancing the risk of misclassification regardless of group membership while the demographic parity metric requires that the probability of favorable outcomes be equal across all demographic distributions.

Introducing these metrics is essential to abide by transparency, accountability and the rights of individuals regarding automated decision-making, as stated by GDPR[1]. With the increasing use of machine learning in decision-making, regulators are also beginning to focus specifically on algorithmic fairness[2]. Understanding and quantifying these metrics allows us to detect potential biases. Once these disparities are identified, fairness mitigation techniques can be applied. This process helps ensure that sensitive groups are not unduly harmed by automated decisions. However, addressing fairness often involves trade-offs with other model performance metrics like precision or overall accuracy. Awareness of these metrics helps in making informed decisions about where and how much to balance these trade-offs. From an ethical perspective, fair treatment of individuals is a core value in responsible data science practices.

Our Standard

For our model, we have opted to use demographic parity and equalized parity metrics. These metrics have been chosen by our intent to balance the equitable distribution of favorable outcomes across groups and to ensure that the model’s error rates are not biased towards a particular section.

Demographic parity ensures that the proportion of favorable outcomes is consistent across groups, promoting a balanced treatment and minimizing disparate impact. Equalized odds, which require similar error rates across groups, are particularly important in maintaining equitable treatment in model performance. Our model demonstrated low scores to begin with, which was slightly increased post-mitigation, showcasing overall fairness preservation. We opted to not apply the recall metric as it is the proportion of actual positives correctly identified by the model. By ensuring Equalized Odds, we are implicitly ensuring that the model's ability to identify positive cases is balanced across groups.

Despite the fairness focused adjustments, the overall accuracy declined only slightly - 0.7467 to 0.7443 - demonstrating improvements in ensuring equity with minimal loss in performance. With metrics increasing slightly post-mitigation and accuracy decreasing slightly, we choose mitigation because prioritizing fairness and reducing disparate impacts is essential, even if it incurs a slight penalty.

Potential Objections & Future Solutions

While our approach to fairness mitigation shows promise, several important objections warrant consideration. First, the accuracy-fairness trade-off (0.7467 to 0.7443) raises questions about the business viability of fairness interventions. To address this, we propose implementing a sliding-scale approach where fairness constraints can be adjusted based on specific use cases and regulatory requirements, potentially using threshold optimization techniques to find optimal balance points as referenced in our class notes.

Some stakeholders might argue that our chosen metrics (Demographic Parity and Equalized Odds) may not fully capture all dimensions of fairness in insurance contexts. For instance, we did not implement recall parity, which could be particularly relevant for ensuring equal opportunity across groups. Future iterations could incorporate a multi-metric evaluation framework that balances different fairness definitions based on specific business and ethical priorities, similar to the approach suggested in the EU AI Act[2].

The fairness metrics we've implemented treat protected attributes as binary categories, potentially oversimplifying the complex, intersectional nature of discrimination. A person's identity encompasses multiple dimensions simultaneously (age, gender, socioeconomic status), and bias may manifest differently at these intersections. We recommend implementing intersectional fairness analysis in future versions to detect and mitigate bias that affects specific subgroups defined by multiple attributes[3].

Critics might contend that our post-processing approach to bias mitigation, while effective, doesn't address underlying data representation issues. To counter this, we propose exploring pre-processing techniques that address biased data representation before model training, potentially through reweighting or sampling methods that ensure balanced representation across protected groups.

Finally, the relatively small improvement in fairness metrics post-mitigation raises questions about whether more aggressive interventions are warranted. We suggest conducting a cost-benefit analysis that quantifies the utility gain for protected groups against overall performance reduction, adopting a utilitarian framework similar to what was discussed in our ethics overview. This would help determine appropriate thresholds for fairness intervention based on measurable impact rather than arbitrary statistical targets.

References:

[1] *Legal text*. General Data Protection Regulation (GDPR). (2024, April 22). https://gdpr-info.eu/

[2] *EU AI act: First regulation on artificial intelligence: Topics: European parliament*. Topics | European Parliament. (n.d.). https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[3] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. Proceedings of the 35th International Conference on Machine Learning.