Towards Developing Fairness-Aware Equitable Models

Amisha Priyadarshini
University of California Irvine
apriyad1@uci.edu

Abstract

We investigate the problem of ensuring fairness and interpretability in machine learning models for high-stakes decision-making, explicitly focusing on student admissions. While data-driven approaches offer scalability and efficiency, they often risk encoding systemic bias into the dataset. These biases might disproportionately impact the admission outcomes of underrepresented groups. Our work seeks to address these challenges by developing fairness-aware deep learning frameworks. To achieve a balance between predictive performance and equitable outcomes, we investigate novel optimization strategies that address the fairness- accuracy trade-off.

1 Introduction

Fairness in Machine Learning (ML) models is not just an ethical requirement but has become a technical challenge. Addressing the bias after deployment is tedious; hence, our research focuses on designing fairness-aware Deep Learning (DL) frameworks that mitigate such biases while maintaining model utility. Our approach aims to solve one of the most pressing challenges in AI for social impact: ensuring that decision-making systems in student admissions are fair and interpretable. Our work is grounded in real-world applicability as we utilize sensitive, real-world datasets provided by the University of California, Irvine (UCI). The UCI dataset from the academic year 2018-19 holds various features, such as student grade point average, advanced placement test scores, participation in educational programs, responses to Personal Insight Questions (PIQs), demographic information, and extra-curricular activities[Priyadarshini *et al.*, 2023]. To uphold the holistic evaluation of the applicant's profile, we consider diverse features to help gain a comprehensive view. However, such real-world datasets exhibit significant structural biases due to long-standing societal iniquities. For instance, representation bias is evident in the under-representation of certain demographic groups. In contrast, historical bias gets embedded due to prior admission outcomes reflecting policies that may have favored a particular demographic group.

Despite the UC Non-discrimination Policy[University of California, Irvine, 2024a], certain important aspects, like socioeconomic status, are not explicitly protected but profoundly impact the admissions process[University of California, Irvine, 2024b]. Ignoring such contextual factors can lead to systemic bias, causing the model to favor applicants from privileged backgrounds implicitly. Our research attempts to address this gap by posing and answering several key questions:

- How do we identify the underlying socioeconomic bias that may be present in our real-world datasets? If present, how does it affect the model's decision-making outcomes?
- What kind of fairness technique implementation in our learning model could efficiently mitigate the bias during the model training without manipulating the datasets?
- Given the inherent fairness-accuracy trade-off, is there a way to balance the model's predictive performance while ensuring it is fair?

2 Our Contributions

2.1 Interpretable Fairness-Aware Deep Learning Framework

The novel adversarial debiasing framework tailored to the admissions domain jointly optimizes for predictive performance and fairness across sensitive attributes. We choose the Input Convex Neural Network (ICNN) architecture trained using the Optimistic Adam (OAdam) optimizer to promote transparency in our models. Due to its property of explicitly capturing the convex relationships between input variables, ICNN ensures the preservation of the convexity property. We use a fully connected ICNN, with k-layers, defined as follows,

$$z_{i+1} = g_i(W_i^{(z)} z_i + W_i^{(y)} y + b_i), f(y, \theta) = z_k$$

where z_i denotes the layer activation, $\theta = W_{i:k-1}^{(z)}$ are the parameters, y is the target variable and g_i are non-linear activation functions. Our work[Priyadarshini *et al.*, 2023] is central to the theorem, which states that the function f is convex in y provided that all $W_{i:k-1}^{(z)}$ are non-negative and all functions g_i are convex and non-decreasing. The guaranteed convexity of the ICNN model ensures a monotonic and interpretable relationship between input features and predictions.

We check the interpretability of the model's predictions by leveraging this property and applying the LIME (Local Interpretable Model-Agnostic Explanations) technique. It analyzes the individual admission decision and provides a transparent, instance-level explanation that aligns with the interpretable nature of our model.

Integrating the Adversarial Debiasing technique into the proposed DL architecture, the classifier model engages in a zero-sum game with the adversary model. Here, the adversary is tasked to predict whether the classifier exhibits unfair behavior concerning the sensitive attributes. Conversely, we propose incorporating SES as a fairness-aware component of the model. On the other hand, the classifier aims to achieve accurate predictions while trying to avoid penalization if the adversary detects unfair decisions. Our method[Priyadarshini and Gago-Masague, 2024] models the loss function of the adversarial debiasing technique following a zero-sum game approach as follows:

$$\min_{\theta_{clf}}\{\text{Loss}(\theta_{clf}) - \lambda \cdot \text{Loss}(\theta_{clf}, \theta_{adv})\}$$

where $\lambda > 0$ is the hyperparameter controlling the strength of fairness, θ_{clf} and θ_{adv} represent the parameters of the classifier and the adversary, respectively. The goal of the loss function is to enhance the classifier's ability to predict the target while reducing its sensitivity concerning the protected attributes.

2.2 Fairness Accuracy Trade-off

For our model training, we utilize the OAdam optimizer due to its property of introducing the optimistic update term, $\Delta\theta_t$, in the parameter update equation. The weight update technique can be described as follows,

$$\theta_t = \theta_{t-1} - 2\eta \cdot s_t + \eta \cdot s_{t-1}$$

where η is the step size, θ_t and θ_{t-1} are the parameters of the model at time step t and $t-1$ respectively. Similarly, s_t and s_{t-1} are the scaled gradients at time step t and $t-1$ respectively. This formulation encourages the model to explore diverse data distribution modes and leads to enhanced training stability. Importantly, OAdam offers a principled solution to handle the fairness-accuracy trade-off. Promoting smoother and more stable convergence dynamics helps the model avoid sharp minima that may favor accuracy at the cost of fairness. This allowed our proposed model to converge to fairer and unbiased solutions more reliably. The proposed framework integrates a fairness parameter that maintains a fairness-accuracy balance, λ. λ can control the trade-off by enabling context-specific fairness calibration during training.

To measure the fairness quotient in our models, we utilize two key fairness metrics: p-$\%$ rule and demographic parity. A high p-$\%$ rule and lower demographic parity scores indicates a fairer model. To holistically assess model quality, we complement fairness metrics with standard performance indicators such as accuracy, precision, recall, F1-score, AUROC. We particularly emphasize the Recall score as it reflects the model's ability to correctly identify qualified applicants. Preferably because a high recall mitigates the risk of disproportionately excluding candidates from underrepresented backgrounds. Further comparing the fully-trained model performance with the pre-trained model based on these fairness metrics provides a clear understanding of bias mitigation. The pre-trained models serve as baseline, where the classifier is trained to predict the target variable, and the adversary learns to predict sensitive attribute. Whereas, the fully-trained models undergo adversarial training that optimizes the classifier to maintain predictive performance while trying to evade the penalization by the adversary. On mapping them onto Kernel Density Estimation (KDE) plots, we see an improvement in the degree of overlapping or similar distribution between the curves in the Fully-trained models for each sensitive attribute. This indicates that each feature has reduced influence over model predictions compared to the Pre-trained models where the degree of overlapping is comparatively lower.

3 Ongoing and Future Work

Our ongoing research aims to address the follwoing key questions:

- Given that the neuro-symbolic models integrate symbolic reasoning with the flexibility of neural networks, they potentially enable better generalization, which leads us to our question: Can we improve the fairness-accuracy trade-off in the adversarial debiasing framework by replacing the traditional neural networks with neuro-symbolic models?
- Could we guarantee fairness in model outcomes under theoretical assumptions about the optimization process and model structure?
- In one of our significant steps, we aim to identify and control for indirect influences of sensitive attributes, enabling the learning model to disentangle correlation from causation. One of our primary goals is to uncover and mitigate hidden data dependencies that contribute to biased outcomes by incorporating causal reasoning.

References

[Priyadarshini and Gago-Masague, 2024] Amisha Priyadarshini and Sergio Gago-Masague. Fair evaluator: An adversarial debiasing-based deep learning framework in student admissions. In *2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 152–161. IEEE, 2024.

[Priyadarshini *et al.*, 2023] Amisha Priyadarshini, Barbara Martinez-Neda, and Sergio Gago-Masague. Admission prediction in undergraduate applications: an interpretable deep learning approach. In *2023 Fifth International Conference on Transdisciplinary AI (TransAI)*, pages 135–140. IEEE, 2023.

[University of California, Irvine, 2024a] University of California, Irvine. Non-discrimination policy. https://grad.uci.edu/non-discrimination-policy/, 2024.

[University of California, Irvine, 2024b] University of California, Irvine. Uci. https://news.uci.edu/2024/07/31/uc-irvine-admits-record-number-of-california-students/, 2024.