Bias Detection And Explainability In Ai Models

Rana Gaber

Abstract

Artificial intelligence models inherit biases from the textual data they are trained on, reflecting human societal biases. This issue is not limited to text; tabular data can also show bias, especially with imbalanced demographic representations. For example, if one demographic is overrepresented, the models may perform poorly for others. In this study, the training dataset was intentionally skewed with more male resumes than female ones to assess the model's performance in the face of such misrepresentation.

1 Dataset description

The dataset includes eleven attributes: Age, Gender, Education level, Years of experience, Number of previous companies, Distance from the company, Interview score, Skill score, Personality score, Recruitment strategy, and Hiring decision. Gender is the only explicitly sensitive attribute, serving as the main axis of potential bias in the study. While other attributes are relevant to candidate evaluation, they are not deemed sensitive and are assumed not to introduce social bias. However, some features may still correlate with sensitive attributes, requiring careful consideration in bias detection and mitigation.

2 Model Architecture

The algorithm employed in this study was a support vector machine (SVM) model. A randomized search methodology was implemented to optimize the choice of the kernel function and the polynomial degree. Ultimately, the evaluation process led to the consideration of two distinct SVM architectures: one utilizing a linear kernel and the other a polynomial kernel of degree two. Given that the performance metrics of both models exhibited minimal variance, the linear SVM was selected for further analysis. This model incorporated a regularization parameter set at **162.975** and demonstrated an accuracy rate of **85.86**.

3 Fairness analysis

To evaluate potential biases in the model, three metrics were used: Demographic Parity, Equal Opportunity, and Average Odds Difference. The Fairlearn package computed the first two metrics, while Average Odds Difference was calculated manually. Prediction rate disparities across gender groups were visualized beforehand, as shown in the accompanying figure.

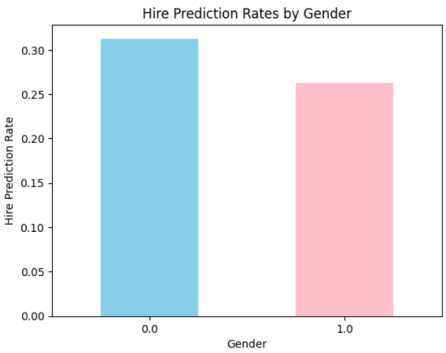

Figure 1: disparities in prediction rates across gender groups

The analysis reveals a discernible, albeit modest, disparity in the model's predictions favoring male applicants over their female counterparts. This observation is likely attributable to the intentional imbalance incorporated into the model's design.

In examining the Disparate Impact Score (DPD), the computed value was **0.024**, indicating that positive decisions were not equitably distributed between genders, resulting in a variance of **2.4** percent. This difference does not raise significant concerns regarding fairness. Conversely,

the Equal Opportunity Difference (EOD) which serves as an indicator of the model's accuracy across gender groups, signifies a substantial disparity in predictive behavior, with a noteworthy difference of 10.8 percent.

Finally, the Average Odds Difference (AOD) was calculated at **0.057** (or 5.7 percent). This metric assesses the mean disparity between the true positive and false positive rates of the two groups. The AOD suggests moderate unfairness in predictive outcomes, though it is less severe than the implications derived from the EOD score alone. This comprehensive analysis underscores the necessity for ongoing evaluation and potential recalibration to enhance equity within predictive models.

4 Explainability results and discussion

to take a closer look at how the model predicts whether an applicant is hired or not, furthermore to see how each attribute affects the classification, two tools were used, which are **SHAP** and **LIME**, where SHAP was used to examine the importance of each attribute generally, and LIME was used to examine five different samples especially.

4.1 SHAP Results

A SHAP explainer has identified the significance of various attributes in the model's decision-making. The Recruitment Strategy is the most influential factor, followed by skill score, educational level, interview score, personality score, and years of experience.

Interestingly, Gender ranks next and shows moderate importance, surpassing distance from the company slightly. However, this raises concerns about potential implicit biases in the model's decision-making regarding Gender.

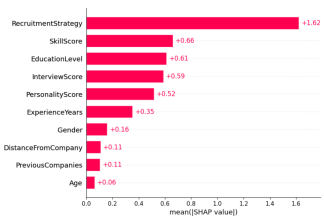

Figure 2: Attribute Importance according to **SHAP**

4.2 LIME Results

upon analyzing the five samples with LIME, it is noticable that the model associates a Gender score that is less than or equal zero to be a good indicator that pushes an applicant a step forward towards being accepted, and unsurprisingly, a Gender score of zero equates to a male applicant, although it was surely not the only factor, going through each sample applicant, you could notice that the following attributes were responsible for getting the applicant hired:

- Sample 1: Recruitment strategy, Personality score, Experience Years, Interview Score, Previous companies, Distance from company and **Gender**
- Sample 2: Recruitment strategy, Experience Years, Interview Score, and Distance from company
- Sample 3:Recruitment strategy, Interview Score, Personality score, Age , Distance from company and previous companies

on the other hand examining the two not-hired applicants, you would notice that the following attributes caused the failure of the the applicants:

- Sample 1: Recruitment strategy, Education level, Interview score, skill score, previous companies, distance from company and **Gender**
- Sample 2: Recruitment strategy, Interview score, skill score, distance from company, **Gender** and Age

which proves that Gender does influence the decision-making process in hiring applicants, albeit to a limited extent.

5 Bias Metigation

To address bias in the model, a correlation remover was applied to eliminate associations with the Gender attribute, using an alpha value of **1.0**. This significant step toward fairness improved the model's accuracy from **85.8** to **87.46**.

Performance metrics showed a slight enhancement in Demographic Parity, increasing by almost **2** to **4.2**. The Equalized Odds difference decreased significantly from **10** to **4.7**, indicating improved fairness. The Average Odds Difference (AOD) also dropped by **3**, resulting in a score of **2.9** instead of **5.7**. These results highlight the effectiveness of the correlation remover in creating a more equitable and accurate model.