Report

Fairness Analysis Report: Gender Classifier

1.Treatment Equality Evaluation

Overview:-

The fairness analysis evaluates how the classification model treats different gender groups by analyzing False Positive Rate (FPR), False Negative Rate (FNR), and the Treatment Equality Ratio (FP/FN). The results provide insights into potential biases in model predictions.

Summary of Results:-

Metric	Man	Woman
True Positives (TP)	227	228
False Positives (FP)	2	5
False Negatives (FN)	5	2
True Negatives (TN)	228	227
False Positive Rate (FPR)	0.0087	0.0216
False Positive Rate (FPR)	0.0216	0.0087
Treatment Equality (FP/FN)	0.40	2.50

Key Observations:-

- **False Positive Rate (FPR):** The FPR for men is significantly lower (0.87%) compared to women (2.16%). This suggests that the model is more likely to incorrectly classify a woman as positive than a man.
- **False Negative Rate (FNR):** The FNR for women (0.87%) is lower than for men (2.16%), indicating that women are less likely to be wrongly classified as negative.
- **Treatment Equality Ratio (FP/FN):** The ratio for men is **0.40**, while for women, it is **2.50**, showing a **significant disparity of 2.10**. This suggests that women experience more false positives per false negative compared to men.

Implications and Fairness Concerns:-

- The model exhibits **disparate treatment** between men and women, especially in how false positives and false negatives are distributed.
- A **higher treatment equality ratio for women** (2.50) means that women are disproportionately subjected to false positives compared to men.

2.Equality of Opportunity Evaluation

Overview

The **Equality of Opportunity** metric measures how well the model provides equal access to positive predictions for different groups. It is quantified using the **True Positive Rate (TPR)**, which represents the proportion of correctly identified positives out of all actual positives.

Summary of Results

Metric	Man	Woman
True Positive Rate (TPR)	0.98	0.99

Key Observations

- **High TPR for both groups:** The model correctly identifies **98% of positive cases for men** and **99% for women**.
- **Minimal disparity:** The **difference in TPR is only 0.01 (or 1%)**, suggesting that the model is nearly equal in its ability to identify positive cases across genders.
- **Near-perfect performance:** Both values indicate that the model performs exceptionally well in detecting positive cases, with **only a minor difference in opportunity distribution.**

Implications and Fairness Concerns

Low disparity in Equal Opportunity: Since the gap between the TPR values is very small (1%), there is **no significant bias** in providing equal access to positive classifications.

References

1. Fairness in Deep Learning: A Survey on Vision and Language Research OTAVIO PARRAGA, MARTIN D. MORE, CHRISTIAN M. OLIVEIRA∗, NATHAN S. GAVENSKI, LUCASS.KUPSSINSKÜ,ADILSONMEDRONHA,LUISV.MOURA,GABRIELS.SIMÕES, and RODRIGO C. BARROS,MachineLearningTheory and Applications (MALTA) Lab, PUCRS, Brazil.

2. A Survey on Bias and Fairness in Machine Learning NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI.

3. Fairness in Machine Learning: A Survey SIMONCATON,University College Dublin, Ireland CHRISTIAN HAAS, University of Nebraska at Omaha, USA and Vienna University of Economics and Business (WU), Austria.