Detecting and Mitigating Bias in Simulated Employment Datasets Using Machine Learning Pipelines

Abstract

Bias in machine learning systems has been widely observed and debated, especially in sensitive domains such as employment, finance, and criminal justice. This paper introduces a simulated dataset with embedded demographic biases to demonstrate the process of identifying and mitigating such biases using machine learning techniques. Using a custom Python pipeline that integrates encoding, bias detection, and mitigation, we analyze the impact of gender and racial bias on salary classification outcomes. The experimental results show that applied bias mitigation strategies can significantly reduce discriminatory correlations, providing a baseline framework for bias-aware machine learning model development.

1. Introduction

The deployment of machine learning (ML) systems in real-world applications poses ethical challenges, particularly when these systems inherit or amplify historical biases. Fairness in AI has become a prominent research topic, with an increasing emphasis on detecting and mitigating algorithmic bias in datasets and models.

This study presents a controlled environment for simulating bias in employment-related data. Using synthetic data generation and customized bias detection and mitigation tools, the paper illustrates how biases can be introduced through demographic distributions and how they affect binary classification tasks such as salary categorization.

2. Related Work

Several works have addressed fairness in machine learning through various frameworks, such as Fairness through Awareness (Dwork et al., 2012), Disparate Impact Remover (Feldman et al., 2015), and Adversarial Debiasing (Zhang et al., 2018). This paper builds upon the practical applicability of bias mitigation tools and contributes a reproducible pipeline for bias experimentation.

3. Methodology

3.1 Dataset Generation:

A sample dataset of 1,000 entries was generated to simulate employee records. The dataset includes the following attributes: age, gender, race, education, experience, and salary. Biased salary assignments were programmatically introduced based on gender, race, and education.

3.2 Data Preprocessing:

Categorical attributes were label-encoded to allow numerical processing.

3.3 Bias Detection:

We employed the BiasDetector class, which computes correlation metrics and bias-related fairness metrics between protected attributes and the target variable.

3.4 Bias Mitigation:

To address detected biases, the BiasMitigator was used. This component applied mitigation strategies (e.g., reweighting, resampling) to adjust the dataset while preserving target performance.

4. Results

4.1 Encoding Overview:

Encoded categorical mappings ensured clarity and traceability.

4.2 Bias Detection Report:

The bias report revealed significant correlation between protected attributes and the salary classification.

4.3 Post-Mitigation Analysis:

After applying mitigation, the correlation between protected attributes and the target dropped significantly. Demographic distributions were normalized, and salary outcomes were more evenly distributed across gender and racial groups.

5. Discussion

Our simulation validates that demographic biases-when present in training data-can influence model predictions. Even in a synthetic setting, biased salary assignments led to skewed outcomes in high salary classification.

The success of the mitigation pipeline demonstrates that automated tools can help reduce such disparities. However, this approach must be supplemented with domain expertise and real-world validation, particularly in high-stakes applications.

6. Conclusion

This study demonstrates a complete machine learning pipeline for bias simulation, detection, and mitigation. The results suggest that systematic bias can be effectively identified and corrected using post-processing strategies. Future work could involve extending the pipeline to real-world datasets and incorporating in-processing techniques such as adversarial training or fairness constraints during model learning.

References

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. In AIES.