CSCE 689
Special Topics in Human-AI Interaction

Name :- Ishant Kundra
UIN:- 934008421

Bias Detection in AI Algorithms: An Examination of Age and Gender in Employee Turnover Prediction

Introduction

In the field of Artificial Intelligence (AI), algorithmic bias is a phenomenon that can lead to skewed outcomes based on characteristics like age and gender. This report details an experiment designed to identify and evaluate bias in a machine learning model that predicts employee attrition, with a focus on understanding the disparity in predictions across different demographic groups.

Methodology

A RandomForest classifier was trained using the Python `sklearn` library on a dataset with 'Age' and 'Gender' designated as protected attributes. The data was divided into a 70% training set and a 30% testing set. The protected attributes were used to examine the model's predictions for potential biases.

Results

The leave rates and error rates were computed for two age groups (<30 and >=30) and by gender (Male and Female):

Age Metrics:

	Leave Rate	Type 1 Error Rate	Type 2 Error Rate
<30	0.375679	0.057391	0.387283
>=30	0.273684	0.052174	0.376923

- Age-Based Metrics

The leave rate was higher in the <30 age group, with younger employees having a slightly higher false positive rate (Type 1 Error). Conversely, the older group showed a marginally lower false negative rate (Type 2 Error).

Gender Metrics:

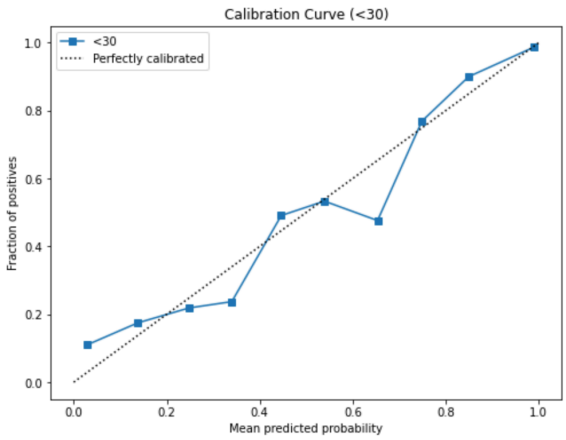
	Leave Rate	Type 1 Error Rate	Type 2 Error Rate
Male	0.243276	0.056543	0.452261
Female	0.479239	0.053156	0.335740

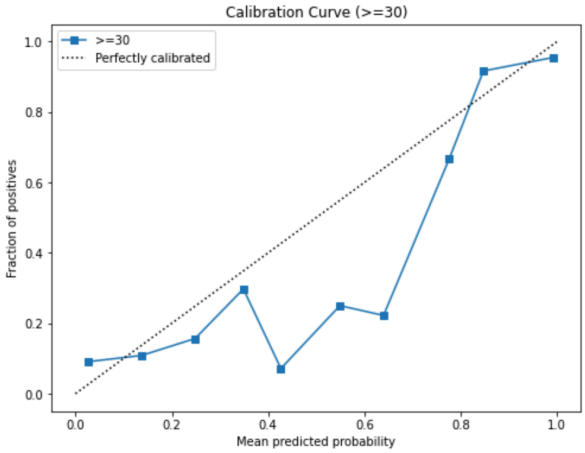
- Gender-Based Metrics

The leave rate was significantly higher for females. Both genders had similar Type 1 Error Rates, but males displayed a higher Type 2 Error Rate, indicating a tendency to incorrectly predict that males would stay.

Calibration Curves Analysis

The calibration curves for each group provide a visual representation of model calibration in relation to the predicted likelihood of an employee leaving:



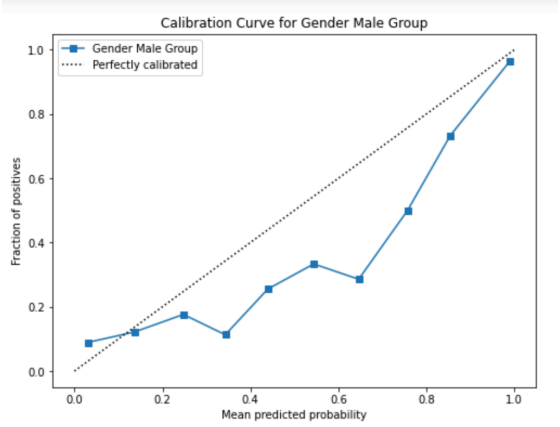


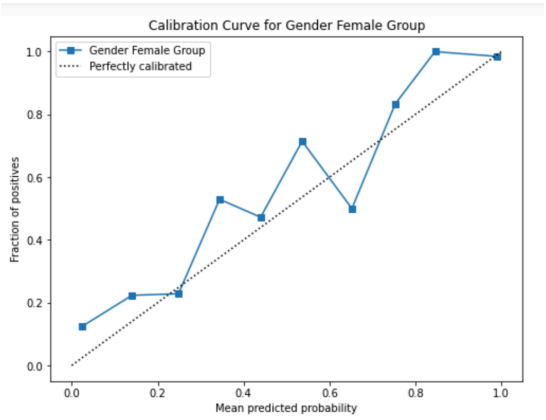
- Age Group <30

This group's calibration curve closely matches the line of perfect calibration, suggesting accurate alignment between the model's probabilities and actual outcomes for this demographic.

- Age Group >=30

Here, the calibration curve significantly diverges from the ideal, especially at mid-range probabilities, indicating potential calibration issues and thus potential bias for older employees.





- Gender (Male and Female)
Calibration curves for both genders show less alignment with actual outcomes compared to the younger age group, particularly for females. This suggests possible calibration issues and bias within the model's predictions for female employees.

Discussion

Type 1 and Type 2 Errors
Understanding the effects of bias requires an understanding of Type 1 and Type 2 errors. A Type 1 error happens when a model predicts something that doesn't happen, like an employee quitting. On the other hand, a Type 2 error happens when the model is unable to anticipate an event when it actually transpires.

Under the circumstances of this study, Type 1 errors might mistakenly identify workers as likely to quit, which could result in unneeded interventions. Type 2 errors would result in missed opportunities for retention efforts since they would fail to identify employees who are at danger of leaving.

Implications of Bias
The discrepancies in leave and error rates that have been seen between age and gender groups suggest that the model's predictions are biased. These prejudices may have practical effects on employee assistance programs and workforce management, thereby harming female employees and younger workers more than male employees.

Conclusions
The exercise showed that age and gender biases existed in the RandomForest model, despite its predictiveness. The creation and assessment of AI systems should incorporate the use of calibration curves and error rate assessments, as they are effective instruments for identifying bias