Assessing Gender Bias in Employee Promotion Prediction: A Machine Learning and Fairness Analysis

NAME: ONAOLAPO Francis Olawale

1. Introduction

Artificial Intelligence (AI) is revolutionizing human resource management by enhancing decision-making in areas like employee promotions. However, when trained on data reflecting historical inequities, AI models can perpetuate biases, particularly against women, embedding unfair patterns into automated decisions (Barocas & Selbst, 2016). For example, if past promotions disproportionately favored men due to systemic barriers, machine learning (ML) systems may disadvantage female candidates, amplifying gender disparities. Chen (2023) highlights the role of partial historical data in the influence of algorithm bias.
This study investigates gender bias in an ML model designed to predict employee promotions, employing fairness metrics to evaluate and mitigate inequities.

Biased AI systems have far-reaching consequences. Employees unfairly denied promotions face economic hardship, diminished job satisfaction, and stunted career growth, with women often disproportionately affected, exacerbating wage gaps. Organizations, meanwhile, risk ethical breaches, legal challenges, and reputational damage, as evidenced by Amazon's scrapped AI hiring tool that penalized women (Dastin, 2018). Research underscores that biased algorithms reinforce systemic inequalities (Mehrabi et al., 2021), while fairness frameworks, such as equal opportunity, offer solutions despite contextual challenges (Hardt et al., 2016). Addressing these issues is critical for fostering equitable workplaces and advancing societal justice.

This study leverages a dataset of 54,808 employee records to develop a Gradient Boosting Classifier (GBC), a robust ML model for promotion prediction. Focusing on gender as a protected attribute, we apply three fairness criteria—equal accuracy (balanced performance across groups), demographic parity (equal selection rates), and equal opportunity (equal true positive rates)—to assess bias. By combining advanced analytics with ethical principles, we aim to illuminate bias in promotion decisions, contributing to fair AI design and promoting workplace equity.

2. Model Development and Fairness Evaluation

2.1 Data Exploration and Preprocessing
The dataset, comprising 54,808 employee records, includes features such as employee identifier, gender, qualifications, department, region, hiring method, training count, age, prior performance rating, service duration, recognition received, and average training performance, with a binary outcome (promoted: 0=no, 1=yes). As shown in figure 1, only 8.5% (4,668) of employees were promoted, indicating class imbalance. Missing values in qualifications (4.4%) and prior performance rating (7.5%) were imputed using mode ("Bachelor's") and median (3.0), respectively. Filling the performance rating with its median (3.0), is a robust choice given its ordinal nature (1–5 scale).

The dataset was examined for duplicate rows, with none identified. Feature checks confirmed no data corruption, ensuring the dataset's integrity and reliability for analysis.

Descriptive statistics revealed a diverse workforce: age (mean=34.8, SD=7.7), service duration (mean=5.9, SD=4.3), and training performance (mean=63.4, SD=13.4).

Figures 1-4 illustrate the distributions of categorical and numerical features. Figure 2 shows the gender distribution across male and female. Figure 5 showed gender disparities in promotions (males: 3,236;

females: 1,432). Figure 5 also depicts the relationship between the target variable and key independent features, revealing an uneven distribution of the target across these features. This imbalance suggests that the independent features may serve as strong predictors of the target variable.

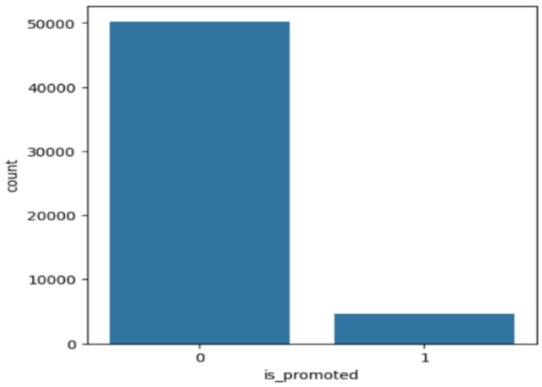


Figure 1: Distribution of the promotion feature.

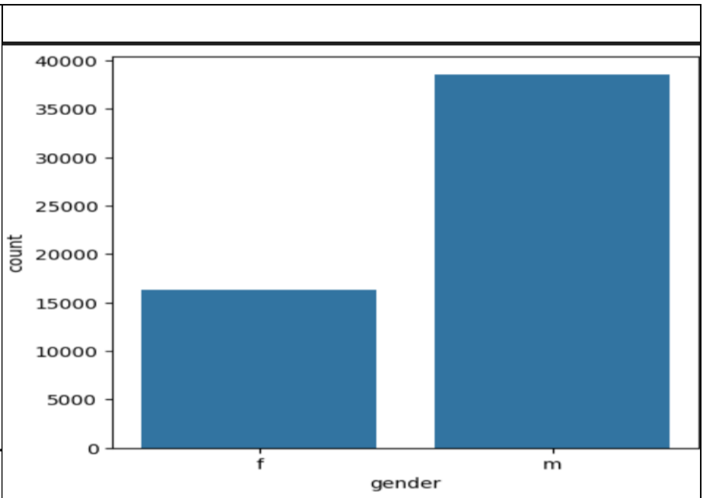


Figure 2: Distribution of the gender feature (the protected feature). 'f' stands for female while 'm' stands for male

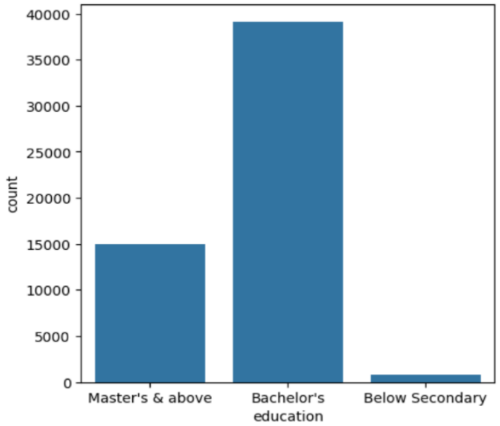


Figure 3: The distribution of the employee education (qualification).

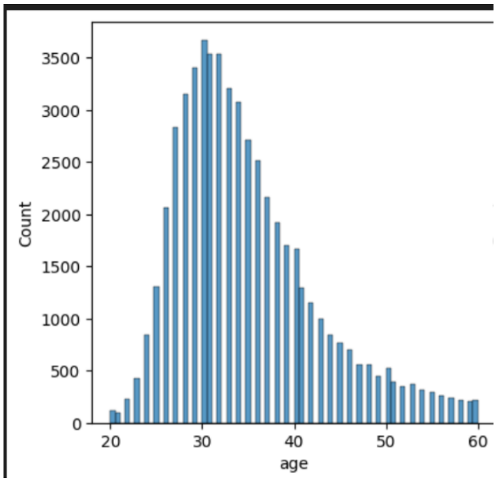


Figure 4: The distribution of the employee age.

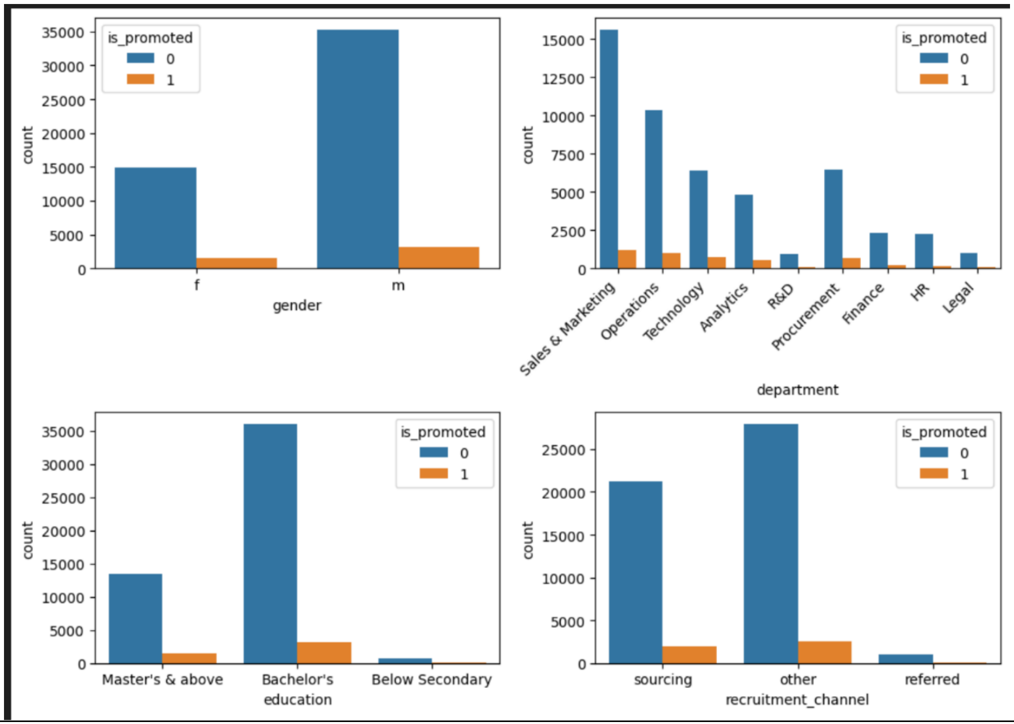


Figure 5: The distribution of the employee promotion feature across the other features.

Figure 6 shows the distribution of the independent features across the gender, which is the assessed protected feature. The distribution of the independent features across the gender are not evenly distributed, which could be because of bias from the dataset. However, the distribution patterns across the male and female gender looks the same, which could be that the dataset is free from bias.

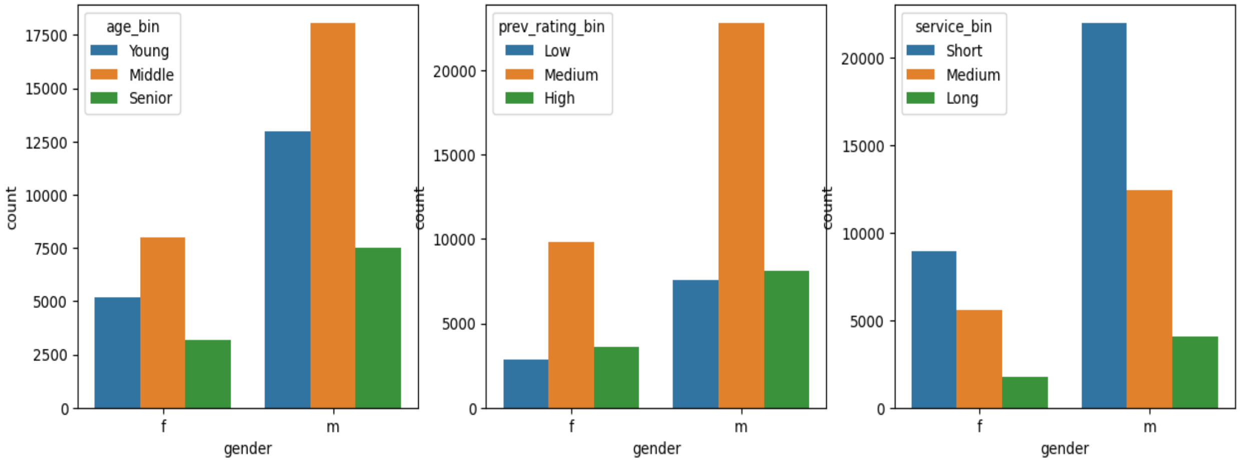
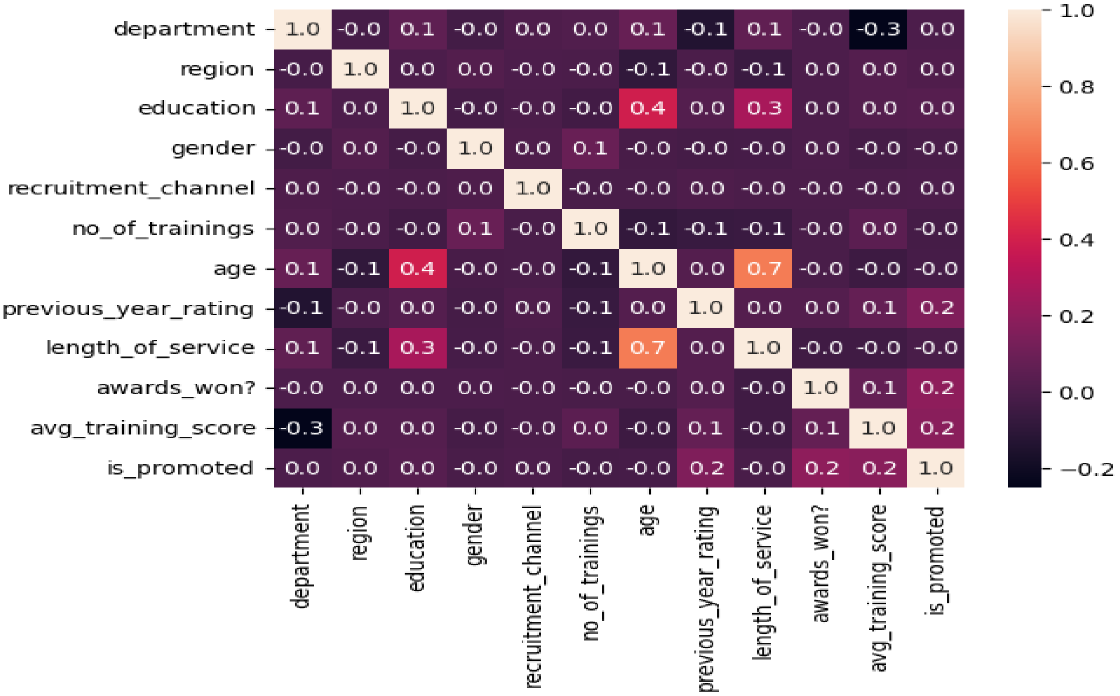


Figure 6: The distribution of the features across the male and female gender (the protected feature).

Categorical features were transformed: gender (male=1, female=0), qualifications (ordinal encoding), and department, region, and hiring method (label encoding). The employee identifier was dropped, yielding 11 predictors.

Figure 7 presents a correlation analysis of the features, revealing weak correlations among education level, age, and length of service. The analysis confirms no multicollinearity, indicating that the features are sufficiently independent for modeling.



	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won?	avg_training_score	is_promoted
department	1.0	-0.0	0.1	-0.0	0.0	0.0	0.1	-0.1	0.1	-0.0	-0.3	0.0
region	-0.0	1.0	0.0	0.0	-0.0	-0.0	-0.1	-0.0	-0.1	0.0	0.0	0.0
education	0.1	0.0	1.0	-0.0	-0.0	-0.0	0.4	0.0	0.3	0.0	0.0	0.0
gender	-0.0	0.0	-0.0	1.0	0.0	0.1	-0.0	-0.0	-0.0	0.0	-0.0	-0.0
recruitment_channel	0.0	-0.0	-0.0	0.0	1.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0
no_of_trainings	0.0	-0.0	-0.0	0.1	-0.0	1.0	-0.1	-0.1	-0.1	-0.0	0.0	-0.0
age	0.1	-0.1	0.4	-0.0	-0.0	-0.1	1.0	0.0	0.7	-0.0	-0.0	-0.0
previous_year_rating	-0.1	-0.0	0.0	-0.0	0.0	-0.1	0.0	1.0	0.0	0.0	0.1	0.2
length_of_service	0.1	-0.1	0.3	-0.0	-0.0	-0.1	0.7	0.0	1.0	-0.0	-0.0	-0.0
awards_won?	-0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	-0.0	1.0	0.1	0.2
avg_training_score	-0.3	0.0	0.0	-0.0	-0.0	0.0	-0.0	0.1	-0.0	0.1	1.0	0.2
is_promoted	0.0	0.0	0.0	-0.0	0.0	-0.0	-0.0	0.2	-0.0	0.2	0.2	1.0

Figure 7: Correlation matrix of the features

2.2 Model Development

The employee promotion model employs a Gradient Boosting Classifier (GBC) due to its robustness with imbalanced datasets and ability to capture complex, non-linear relationships. GBC iteratively constructs decision trees, optimizing log-loss for classification, making it suitable for the 8.5% positive class prevalence. The dataset, comprising 54,808 records, was split into 80% training (43,846 records) and 20% testing (10,962 records), stratified to maintain class proportions. The GBC was configured with 100 estimators, a maximum depth of 3, and a learning rate of 0.1, trained on 11 features. Feature scaling was unnecessary due to GBC's robustness to unscaled data.

Label encoding, data splitting, and the Gradient Boosting Classifier (GBC) algorithm are imported from the scikit-learn library, enabling efficient data preprocessing and model training.

Feature importances, illustrated in Figure 8, reveal the relative impact of each feature on predictions. Notably, the gender feature exhibits minimal influence, indicating a low risk of gender-based bias in the model's decisions. This supports the model's alignment with ethical AI standards.

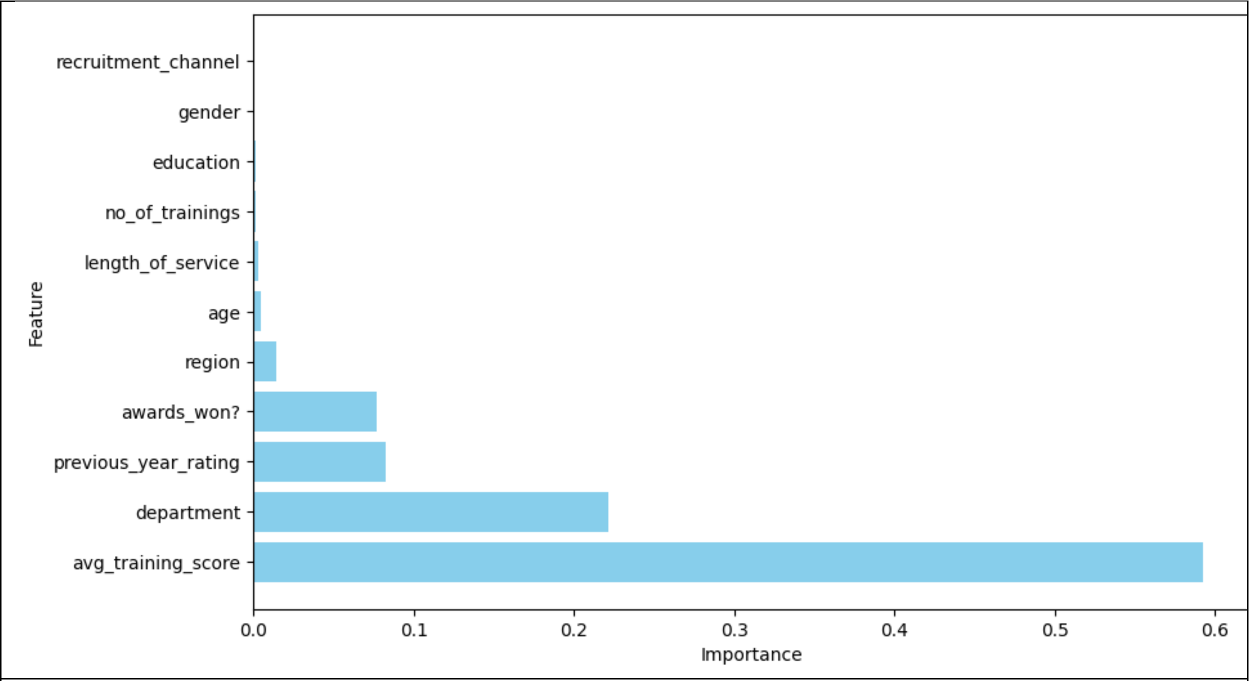
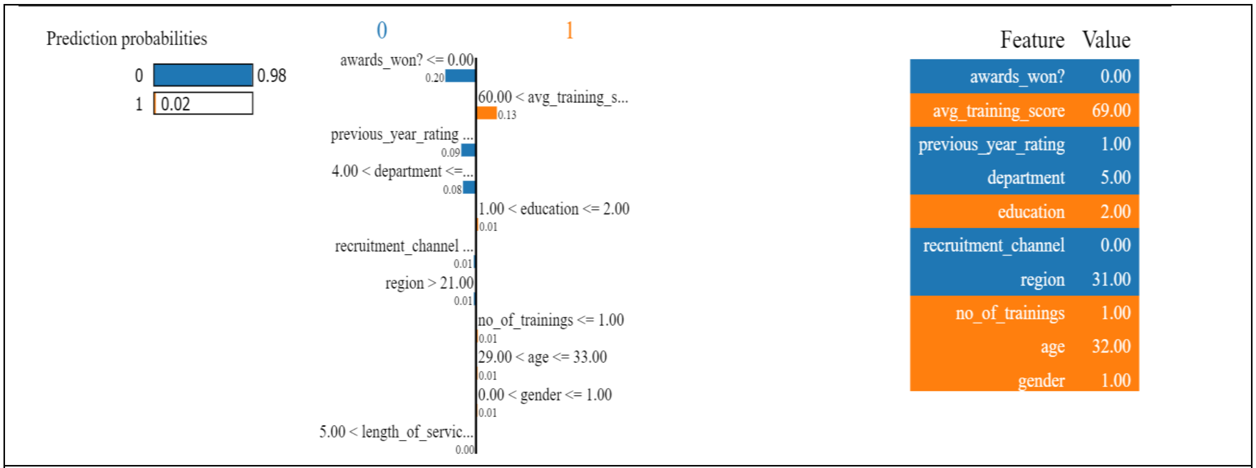


Figure 8: The bar chart of the model feature importance.

To enhance transparency and facilitate bias auditing, Local Interpretable Model-agnostic Explanations (LIME) was implemented. LIME provides clear, model-agnostic explanations for individual predictions, as shown in Figure 9. It highlights that high training performance and recognition are strong predictors of promotions, reinforcing interpretability. It also highlights gender, along with the service duration (length of service) as weak predictors of promotions. This indicates a low risk of gender-based bias in the model's decisions for this dataset instance. By illuminating the factors driving specific predictions, LIME ensures the model adheres to ethical AI principles, enabling stakeholders to audit and trust the decision-making process. This approach underscores the commitment to fairness and accountability in the promotion model's development.



Feature	Value
awards_won?	0.00
avg_training_score	69.00
previous_year_rating	1.00
department	5.00
education	2.00
recruitment_channel	0.00
region	31.00
no_of_trainings	1.00
age	32.00
gender	1.00

Figure 9: The LIME Local Explanation for one of the dataset instance

2.3 Performance Evaluation

The model performance was evaluated with the use of confusion matrix. Figure 10 shows the confusion matrix of the model predictions for the test set, where the True Negatives (TN) is 10,015; False Positives (FP) is 13; False Negatives (FN) is 647; True Positives (TP) is 287

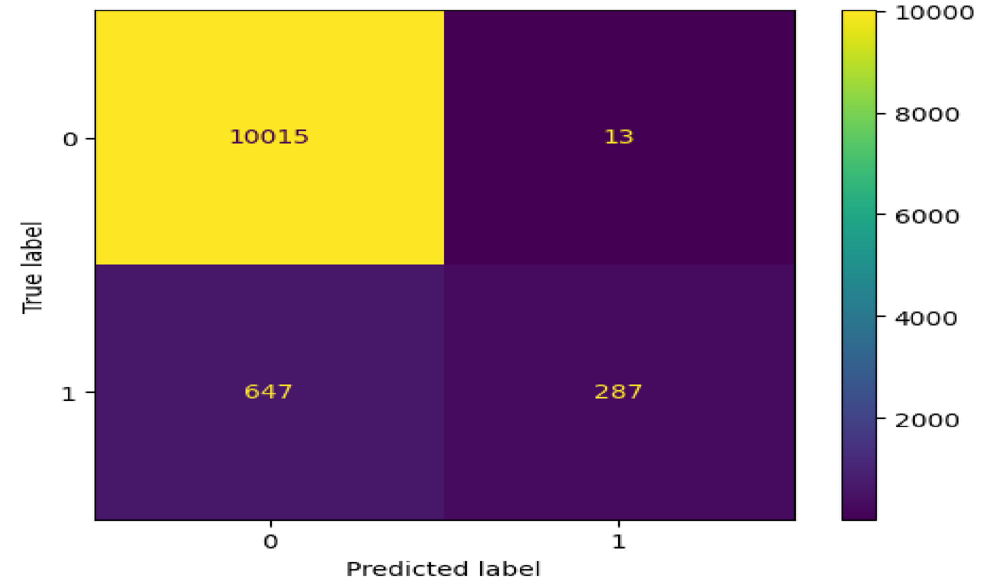


Figure 10: The model predictions confusion matrix

Model performance was thoroughly assessed using training and test dataset results. Comparing training and test metrics ensured detection of overfitting or underfitting, validating the model's generalization and reliability.

The training set performance metrics were:

- Accuracy: ((TP+TN) / (TP+TN+FP+FN)) = 0.940
- Precision: ((TP) / (TP+FP)) = 0.958
- Recall: ((TP) / (TP+FN)) = 0.314

The test set performance metrics were:

- Accuracy: 0.940
- Precision: 0.957
- Recall: 0.307
- False Positive Rate: ((FP) / (FP+TN)) = 0.001

The model achieves high training and test accuracies (both 94%), indicating strong performance with minimal bias. Nearly identical training and test metrics, with precision differing by 0.001 and recall by

0.007, demonstrate low variance. This combination of low bias and low variance reflects a stable, well-generalized model, critical for effective machine learning development.

The high accuracy (0.940) indicates strong overall performance, and the precision (0.957) reflects reliable positive predictions—important for promotion decisions—while recall (0.307) highlights a limitation: only 30% of actual promotions were identified. This aligns with the class imbalance, as the model prioritizes the majority class (non-promoted), a common challenge in imbalanced datasets.

The low false positive rate (0.001) is considered good for using the model for promotion prediction, as it will ensure the number of people promoted that are not supposed to be promoted is kept minimal.

2.4 Fairness Criteria Application

To evaluate gender bias in the GBC model's promotion predictions, the test set was split by gender (7,748 males, 3,214 females) and assessed using three fairness criteria: equal accuracy (balanced prediction accuracy), demographic parity (equal positive prediction rates), and equal opportunity (equal true positive rates). These metrics determine whether the model treats male and female employees equitably, addressing potential disparities in automated promotion decisions.

Figures 11 and 12 present the confusion matrices for male and female predictions, respectively. For males, the matrix shows 7,086 true negatives (TN), 10 false positives (FP), 449 false negatives (FN), and 203 true positives (TP). For females, the matrix reports 2,929 TN, 3 FP, 198 FN, and 84 TP. These values highlight the model's performance, with false negatives indicating missed promotions, which may disproportionately impact women due to their underrepresentation.

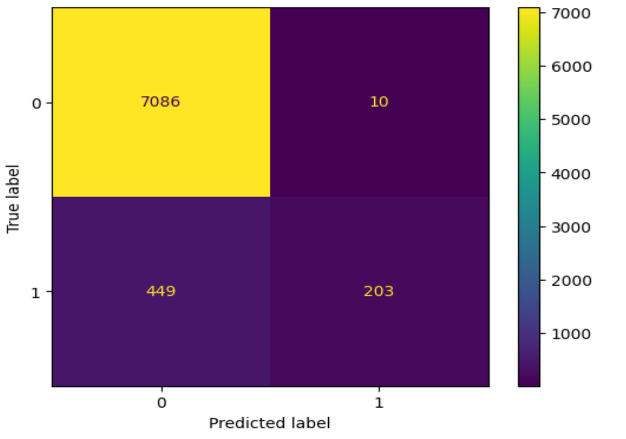


Figure 11: The male predictions confusion matrix

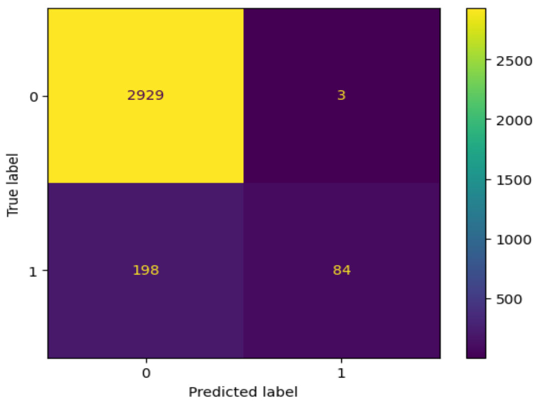


Figure 12: The female predictions confusion matrix

Applying the fairness criteria yields the following results:

- Equal Accuracy: Male accuracy is 0.941, female accuracy is 0.938 (difference: 0.003), indicating nearly identical prediction accuracy across genders.
- Demographic Parity: Male positive rate is 0.028, female positive rate is 0.027 (difference: 0.001), showing consistent selection rates.

- Equal Opportunity: Male recall is 0.311, female recall is 0.298 (difference: 0.013), reflecting a slight male-favored disparity.

The close alignment across metrics suggests minimal gender bias, though the 0.013 equal opportunity gap, likely tied to historical promotion imbalances (652 male vs. 282 female promotions), warrants attention. This minor disparity underscores the need for ongoing fairness monitoring to ensure equitable promotion outcomes, particularly for underrepresented female employees, fostering a more inclusive workplace.

3. Findings

The fairness analysis of the Gradient Boosting Classifier (GBC) model reveals equitable performance across genders, with minimal disparities in promotion predictions. The equal accuracy metric, measuring accuracy rates across genders, shows a difference of 0.003, indicating balanced performance across male and female employees. Demographic parity, measuring equal selection rates, has a difference of 0.001, reflecting consistent prediction rates. Equal opportunity, assessing equal true positive rates, shows a slight disparity (difference of 0.013), likely due to historical imbalances (652 male promotions versus 282 female promotions). With women underrepresented (3,214 vs. 7,748 males), this disparity, though small, suggests potential impacts on female career progression. Missed promotions (false negatives: 449 males versus 198 females) may disproportionately affect women, slowing gender equity, though the model's fairness mitigates severe consequences.

Limitations of fairness criteria are significant. Equal accuracy may obscure class-specific disparities, while demographic parity might enforce artificial equity if merit differs across groups. Equal opportunity overlooks false positives (10 males vs. 3 females), potentially favoring less-qualified men. Historical data imbalances and unmodeled factors, such as workplace culture, may subtly skew predictions, reflecting societal biases rather than individual merit (Mehrabi et al., 2021). These limitations highlight the complexity of achieving true fairness in AI systems (Hardt et al., 2016).

To enhance fairness, strategies like oversampling female records or adopting fairness-aware algorithms could reduce disparities. Continuous monitoring and inclusion of contextual features (e.g., mentorship access) are critical to ensure equitable outcomes. This study underscores the need for comprehensive AI fairness approaches, balancing technical rigor with ethical considerations to foster inclusive workplaces and advance gender equity in promotion decisions.

References

Barocas, S. and Selbst, A.D. (2016) ‘Big Data’s Disparate Impact’, *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.2477899.

Chen, Z. (2023) ‘Ethics and discrimination in artificial intelligence-enabled recruitment practices’, *Humanities and Social Sciences Communications*, 10(1), p. 567. Available at: https://doi.org/10.1057/s41599-023-02079-x.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Available at: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG (Accessed: 25 March 2025).

Hardt, M., Price, E. and Srebro, N. (2016) ‘Equality of Opportunity in Supervised Learning’. arXiv. Available at: https://doi.org/10.48550/arXiv.1610.02413.

Mehrabi, N. *et al.* (2022) ‘A Survey on Bias and Fairness in Machine Learning’. arXiv. Available at: https://doi.org/10.48550/arXiv.1908.09635.