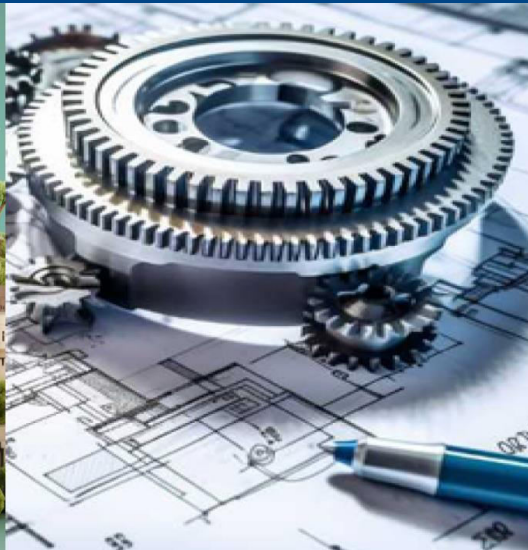
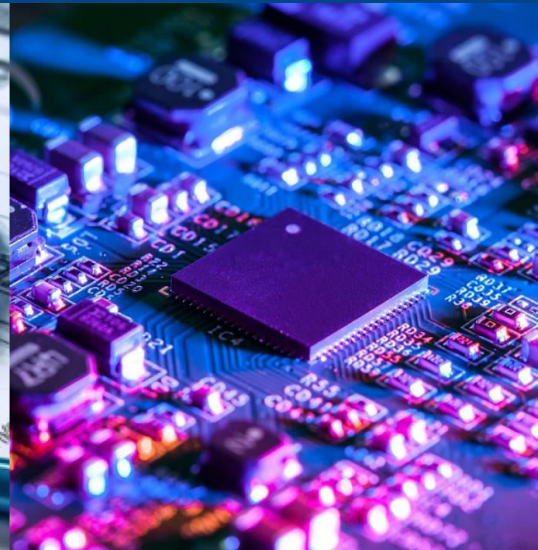
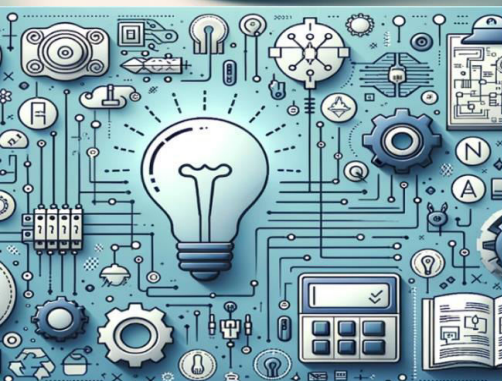




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.864

Volume 9, Issue 10, October 2026



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Algorithmic Fairness in Psychometric Hiring Assessments: An Empirical Audit and Mitigation Framework

A Fairness–Utility Analysis of Behavioral Role-Fit Prediction on the IBM HR Analytics Dataset

Purushothamman Elango Pillai, Dr. Mohammad Abuzar Ansari

MSc, Postgraduate Student, Department of Data Science, SIES College of Arts, Science, and Commerce (Empowered Autonomous), Sion (West), Mumbai, India

HOD, Department of Data Science, SIES College of Arts, Science, and Commerce (Empowered Autonomous), Sion (West), Mumbai, India

ABSTRACT: AI-assisted behavioral and psychometric assessments are increasingly proposed to support hiring and workforce decisions, yet models trained on historical organizational data can reproduce demographic disparities. This paper presents an empirical fairness audit and mitigation study for a role-fit prediction task. Employee attrition in the IBM HR Analytics dataset (1,470 records) serves as a proxy outcome, and satisfaction, involvement and work-context fields stand in for behavioral assessment inputs. Logistic Regression and Random Forest classifiers were trained with class-imbalance handling and audited with Fairlearn across gender, age group and marital status, using demographic parity difference, equalized odds difference and subgroup recall. Age group showed the largest disparity: the equalized odds difference was 0.43 for Logistic Regression and 0.49 for Random Forest, and recall for employees aged 45 and above was 33.3 percent, against 70.6 to 82.4 percent for employees under 30. Two mitigation methods were then applied to age group, a post-processing method (ThresholdOptimizer) and a training-time reduction method (ExponentiatedGradient). Every mitigation narrowed the recall gap, but all except one lowered recall for the 45-and-above group, in two cases to zero, a pattern known as levelling down. Only Logistic Regression with ExponentiatedGradient narrowed the gap (0.43 to 0.11) while raising recall for the 45-and-above group (33.3 to 55.6 percent). The results come from a single stratified split in which that subgroup contains few positive cases, so they are indicative rather than conclusive. The study shows that a fairness metric should be reported together with the recall of the group it is meant to protect.

KEYWORDS: Algorithmic fairness; bias mitigation; employee attrition prediction; psychometric assessment; equalized odds; Fairlearn; levelling down.

I. INTRODUCTION

Organizations increasingly use data-driven tools to screen candidates and manage employees, including assessments that score behavioral and psychological characteristics such as stress tolerance, teamwork orientation and adaptability. These tools learn from historical records, and historical records can carry past inequities. A widely reported example is an experimental recruiting engine at Amazon that, according to Reuters, learned to penalize résumés containing the word “women’s” [1]. A study of vendors of algorithmic pre-employment assessments examined how such assessments are built, validated and checked for bias, and reported that vendors’ formulations of bias, and their ways of dealing with it, differ [2]. Independent empirical audits are therefore needed.

Most published work on employee attrition prediction reports accuracy and feature importance. One analysis of the IBM HR Analytics dataset applies Random Forest, k-means clustering and logistic regression to identify the drivers of attrition and does not examine group-level fairness [3]. Fairness studies on the same dataset have centred on gender:



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Liu et al. used it to model gender bias in staff promotions with Bayesian networks [4], and a survey of datasets used in fairness research lists it with gender as the sensitive feature [5]. Within the sources reviewed for this study, none audits attrition predictions on this dataset across several demographic attributes and compares a post-processing and a training-time mitigation method on both fairness metrics and subgroup recall.

This paper addresses that gap. It treats attrition as a proxy for behavioral role fit, trains two classifiers, audits them across gender, age group and marital status, and applies two mitigation methods to the attribute with the largest disparity. Its central point is that a mitigation should be judged on two things together: the fairness metric, and the recall obtained by the group the mitigation is meant to protect.

A. Problem Statement

AI-based psychometric and behavioral assessment tools are used to judge how well a person fits a role, offering more consistency and scale than unstructured interviews. Because they learn from historical organizational data, they can inherit disparities linked to gender, age, marital status and other characteristics. A model that misses at-risk employees in one age group more often than in another, for example, would give that group less retention support even if its overall accuracy looks acceptable.

Existing work on attrition and hiring prediction tends to stop at accuracy, explainability or a single fairness measurement. What is missing is a complete audit-and-mitigation cycle: measuring bias across several attributes, applying more than one kind of mitigation, and checking whether the mitigation improves outcomes for the disadvantaged group rather than only narrowing a metric.

B. Research Objectives

To address this problem, this research pursues five objectives:

1. **Model Development:** To train Logistic Regression and Random Forest classifiers that predict employee attrition as a proxy for behavioral role fit, using satisfaction, involvement and work-context fields as stand-ins for behavioral assessment inputs.
2. **Empirical Fairness Audit:** To audit these models across gender, age group and marital status using demographic parity difference, equalized odds difference and subgroup recall.
3. **Mitigation Comparison:** To apply a post-processing method (ThresholdOptimizer) and a training-time method (ExponentiatedGradient) to the attribute with the largest disparity and compare their effects.
4. **Fairness–Utility Analysis:** To identify which combinations of model and mitigation narrow the disparity without lowering recall for the disadvantaged group.
5. **Practical Recommendations:** To draw recommendations for the design and evaluation of AI-based psychometric hiring assessment tools from the findings.

II. LITERATURE REVIEW

A. Bias in Automated Hiring and Psychometric Assessment

Concerns about automated hiring appear in both industry reporting and research [1], [2]. Raji and Buolamwini described the audit design and disclosure procedure of the Gender Shades study of commercial facial-analysis products and examined the commercial response to publicly naming biased results [6]. Their work shows how an external audit can bring group-level disparities to light, which is the approach followed here for a hiring-assessment setting. Surveys and textbooks organise the sources of bias and the competing definitions of fairness [7], [8].

B. Fairness Metrics and the Levelling-Down Problem

Group fairness criteria compare a model's behavior across protected groups. Demographic parity compares the rate of positive predictions. Hardt, Price and Srebro proposed equalized odds and equal opportunity, which compare true-positive and false-positive rates [9]. Individual fairness, introduced by Dwork et al., instead asks that similar individuals be treated similarly [10]. Because these criteria can disagree, this study reports both demographic parity difference and equalized odds difference, together with subgroup recall.

Mittelstadt, Wachter and Russell observe that many fairness measures suffer from levelling down: parity is reached by making every group worse off, or by pulling better-served groups down to the level of the worst-served one, and the



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

field rarely measures or reports enough to justify this [11]. That concern motivates reading every fairness result in this paper next to the recall of the group being protected.

C. Mitigation Methods and Imbalanced Data

Fairlearn is an open-source toolkit with metrics for assessing a model's output across groups and algorithms for mitigating unfairness [12], [13]. Its ExponentiatedGradient method follows the reductions approach of Agarwal et al., which reduces fair classification to a sequence of cost-sensitive classification problems [14]. Its ThresholdOptimizer is a post-processing method, in the spirit of the threshold adjustment of Hardt et al. [9], that chooses group-specific decision thresholds on the scores of a trained model. Because attrition is a minority outcome, imbalance handling is also needed; SMOTE creates synthetic minority-class examples instead of duplicating existing ones [15].

D. Prior Work on the IBM HR Analytics Dataset and Research Gap

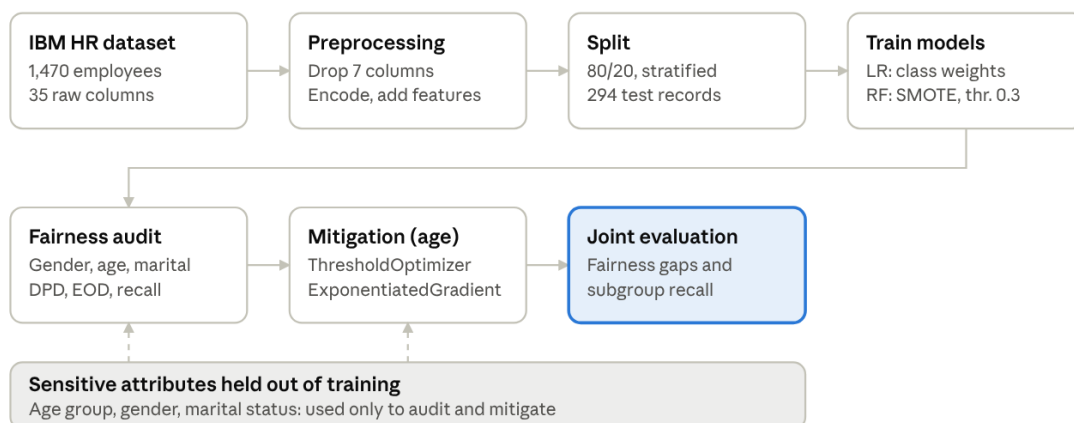
The IBM HR Analytics dataset is a fictional dataset created by IBM data scientists [5], [16]. Yang and Islam applied Random Forest, k-means clustering and logistic regression to it, removed the same constant, identifier and rate columns removed in this study, and reported average test accuracies of 0.8456 for Random Forest and 0.8843 for logistic regression [3]. Their findings also associate age and marital status with attrition, which is the pattern that raises fairness questions here. Liu et al. used the dataset to demonstrate a Bayesian-network method for fair equality of opportunity, with gender as the sensitive variable, education as a proxy for talent and recent promotion as the target [4].

The studies reviewed do not report a combined audit across age, gender and marital status, nor a comparison of mitigation methods in terms of subgroup recall. This research addresses that combination for attrition prediction.

III. METHODOLOGY

The study follows a five-phase pipeline (Fig. 1): dataset framing, preprocessing, model training with imbalance handling, fairness auditing, and bias mitigation. Sensitive attributes are held out of the models' inputs and used only for auditing and mitigation.

Each model is audited, mitigated, then judged on fairness and recall



pipeline · 5 phases, 7 steps, sensitive attributes held out

Fig. 1. Pipeline of the study. LR: Logistic Regression; RF: Random Forest; DPD: demographic parity difference; EOD: equalized odds difference.

A. Phase 1: Dataset and Problem Framing

The study uses the IBM HR Analytics Employee Attrition and Performance dataset [16]: 1,470 employee records and 35 raw columns, with no missing values. Attrition (Yes or No) is the target. Of the 1,470 employees, 237 (16.1 percent) left and 1,233 stayed. No dataset was available that links psychometric assessment scores to role-fit outcomes and



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

demographic fields, so attrition serves as a proxy for poor role fit, on the premise that a poor match between person and role contributes to departure. This is a modeling assumption, not a measurement of psychological fit. Five satisfaction and involvement fields (JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction, WorkLifeBalance and JobInvolvement) stand in for behavioral assessment inputs, alongside job and tenure fields. The dataset is fictional [5], [16], so the findings illustrate a method rather than conditions at a real employer.

B. Phase 2: Preprocessing and Feature Engineering

Seven columns were removed: EmployeeCount, StandardHours and Over18 (constant values), EmployeeNumber (an identifier), and DailyRate, HourlyRate and MonthlyRate, which were nearly uncorrelated with each other and with MonthlyIncome, as also noted in earlier work [3]. Attrition, OverTime and Gender were label-encoded. BusinessTravel, Department, EducationField, JobRole and MaritalStatus were one-hot encoded with the first level dropped. PerformanceRating, which takes only the values 3 and 4, was left out of the model inputs to keep an outcome-type variable out of the predictors.

Two features were engineered. The composite satisfaction score is the mean of four satisfaction-related fields (Eq. 1), and the tenure ratio expresses time in the current role relative to time at the company (Eq. 2). Age was also bucketed into three groups (under 30, 30 to 45, 45 and above) for the fairness analysis.

$$\text{AvgSatisfaction} = \frac{1}{4} (\text{EnvSat} + \text{JobSat} + \text{RelSat} + \text{WLB}) \tag{1}$$

$$\text{TenureRatio} = \frac{\text{YearsInCurrentRole}}{\text{YearsAtCompany} + 1} \tag{2}$$

Gender, age, age group and marital status, including the encoded indicators derived from them, were separated from the feature matrix and kept only for auditing and mitigation. A column derived from a sensitive attribute is itself treated as sensitive, so the age group was excluded from the predictors together with age. Other predictors can still correlate with these attributes, so excluding them does not remove all indirect influence.

C. Phase 3: Class-Imbalance Handling and Model Training

The data were split 80/20 with stratification on the target, giving 1,176 training and 294 test records, 47 of the latter being employees who left. Features were standardised for Logistic Regression and left unscaled for Random Forest. Models were built with scikit-learn [17] and imbalanced-learn [18]. Trained without imbalance handling, both models reached 84 to 86 percent accuracy but flagged few leavers (Table II). Balanced class weights raised the recall of Logistic Regression. For Random Forest, class weights left recall unchanged, so SMOTE [15] was applied to the training data only, and the decision threshold was lowered from 0.5 to 0.3 (Eq. 3). The value 0.3 was chosen after seeing that 0.5 gave low recall; no separate validation set was used for this choice, which is discussed in Section VI.

$$\hat{y}(x) = \mathbf{1} \left[\hat{p}(x) \geq \tau \right], \quad \tau = 0.3 \tag{3}$$

D. Phase 4: Fairness Auditing

Fairness was assessed on the test set for three sensitive attributes: gender (a binary indicator), age group (three levels) and marital status (represented by a married versus not-married indicator). For each attribute, Fairlearn's MetricFrame gives subgroup accuracy and recall, and two summary measures are computed. The demographic parity difference (Eq. 4) is the gap in positive-prediction rates between the groups with the highest and lowest rates. The equalized odds difference (Eq. 6) is the larger of the largest gap in true-positive rate and the largest gap in false-positive rate, where the true-positive rate of group a is the recall of that group (Eq. 5).

$$\text{DPD} = \max_a \Pr(\hat{Y}=1 \mid A=a) - \min_a \Pr(\hat{Y}=1 \mid A=a) \tag{4}$$

$$\text{TPR}_a = \Pr(\hat{Y}=1 \mid Y=1, A=a) \tag{5}$$

$$\text{EOD} = \max \left(\max_a \text{TPR}_a - \min_a \text{TPR}_a, \max_a \text{FPR}_a - \min_a \text{FPR}_a \right) \tag{6}$$

Recall is the share of employees who left that the model flagged. A group with low recall is under-served by an attrition-risk tool, so recall by group is read alongside the two summary measures.

E. Phase 5: Bias Mitigation

Two mitigation methods were applied to age group, the attribute with the largest disparity. ThresholdOptimizer is a post-processing method: given a trained model's scores, it chooses group-specific thresholds to satisfy a fairness



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

constraint, here equalized odds. ExponentiatedGradient is a training-time reduction method: it retrains the base learner several times with changing sample weights so that the resulting ensemble satisfies the constraint, here equalized odds. Both used Fairlearn defaults apart from the constraint and the use of predicted probabilities. Mitigated models were evaluated with the same metrics as the baselines, so that fairness and recall can be compared before and after.

IV. EXPERIMENTAL SETUP

All experiments were run in a Python notebook environment (Google Colab) with pandas, scikit-learn, imbalanced-learn and Fairlearn. A single random seed (42) was used for the split and for the models. Six configurations were compared (Table I). Performance was measured with accuracy, precision, recall and F1-score, and fairness as described in Section III-D. The mitigated models were evaluated by subgroup accuracy and recall by age group and by the two fairness differences.

Table I. Model and mitigation configurations compared. Mitigation uses age group as the sensitive feature; other settings are library defaults.

Configuration	Model and training data	Decision rule
LR	Logistic Regression, balanced class weights, scaled features, original training data	Probability ≥ 0.5
LR + ThresholdOptimizer	Same learner, passed to ThresholdOptimizer, fitted on the original training data	Group-specific thresholds, equalized odds
LR + ExponentiatedGradient	Logistic Regression with balanced class weights, retrained under the equalized odds constraint on the original training data	Ensemble prediction of the reduction
RF	Random Forest (100 trees), trained on SMOTE-resampled data	Probability ≥ 0.3
RF + ThresholdOptimizer	Random Forest (100 trees), passed to ThresholdOptimizer, fitted on the original training data	Group-specific thresholds, equalized odds
RF + ExponentiatedGradient	Random Forest (100 trees, balanced class weights), retrained under the equalized odds constraint on the original training data, without SMOTE	Ensemble prediction of the reduction

The Random Forest configurations are not trained identically: the baseline uses SMOTE and a 0.3 threshold, whereas the mitigated versions were fitted on the original training data (Section VI).

V. RESULTS AND DISCUSSION

A. Exploratory Findings

Overall, 237 of 1,470 employees (16.1 percent) left. Attrition differs most by age and marital status and little by gender (Fig. 2). Employees under 30 leave at more than twice the rate of older employees, and single employees at about twice the rate of married employees or more. Among work-context variables, overtime showed the widest gap, 30.5 percent with overtime against 10.4 percent without. These raw differences are why age and marital status deserve auditing; they are associations, not causes.

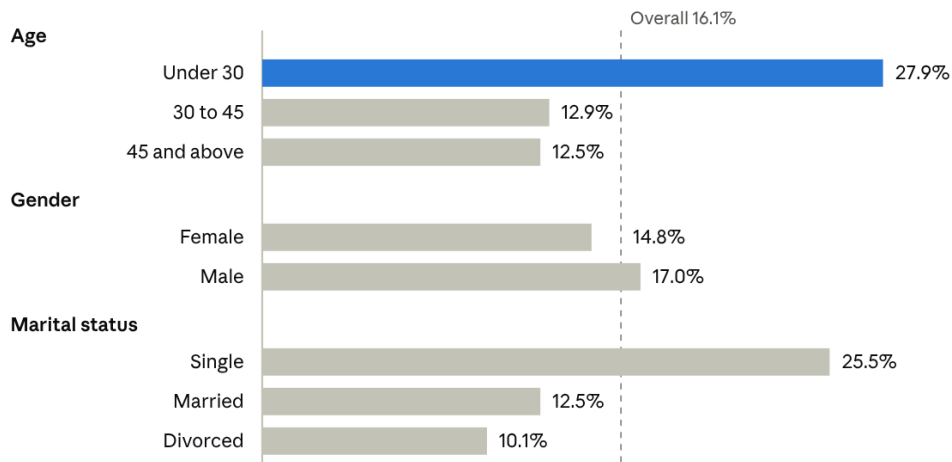


International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Employees under 30 leave at more than twice the rate of older groups

Attrition rate by subgroup, percent of employees; dashed line marks the overall rate



computed from the 1,470-record IBM HR Analytics file (WA_Fn-UseC_-HR-Employee-Attrition.csv)

Fig. 2. Attrition rate by age group, gender and marital status in the IBM HR Analytics dataset (n = 1,470).

B. Baseline Predictive Performance

Table II. Test-set performance of Logistic Regression (LR) and Random Forest (RF) at each imbalance-handling step (n = 294, 47 positives). The two rows marked “audited” are the models used in the fairness analysis.

Configuration	Accuracy	Precision	Recall	F1-score
LR, no imbalance handling	0.8571	0.5926	0.3404	0.4324
RF, no imbalance handling	0.8401	0.5000	0.1277	0.2034
LR, balanced class weights (audited)	0.7551	0.3563	0.6596	0.4627
RF, balanced class weights	0.8537	0.7500	0.1277	0.2182
RF, SMOTE, threshold 0.5	0.8265	0.4333	0.2766	0.3377
RF, SMOTE, threshold 0.3 (audited)	0.7381	0.3438	0.7021	0.4615

Accuracy alone is misleading here. Because 84 percent of employees stay, a model can score above 84 percent while flagging few leavers, as both unadjusted models did. The two audited models reach similar F1-scores (0.463 and 0.462) and recall (0.66 and 0.70) at the cost of accuracy and precision. Their confusion matrices were TN = 191, FP = 56, FN = 16, TP = 31 for Logistic Regression and TN = 184, FP = 63, FN = 14, TP = 33 for Random Forest.

C. Fairness Audit

Table III. Demographic parity difference (DPD) and equalized odds difference (EOD) of the audited models, by sensitive attribute (test set).

Sensitive attribute	LR DPD	LR EOD	RF DPD	RF EOD
Gender	0.0665	0.0895	0.0017	0.1169
Age group	0.1973	0.4286	0.2654	0.4902
Marital status (married vs. not married)	0.0736	0.0785	0.0745	0.2375

Age group shows the largest disparity under both models and both measures. Its equalized odds difference (0.4286 for LR, 0.4902 for RF) is roughly four to five times that of gender (0.0895 and 0.1169). The source is recall: employees aged 45 and above who left were flagged at a rate of 0.33 by both models, against 0.76 for the 30 to 45 group and 0.71



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

(LR) or 0.82 (RF) for employees under 30. The 45-and-above test subgroup contains nine employees who left, so each missed case moves its recall by about eleven points (Section VI).

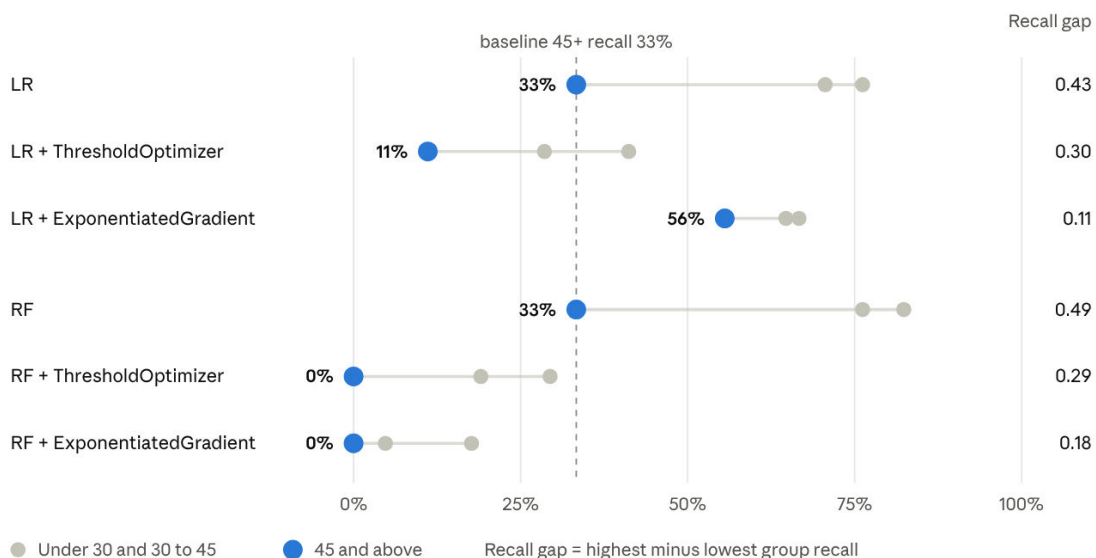
Gender and marital status show smaller gaps. Recall was 0.69 for women and 0.65 for men under LR, and 0.63 and 0.74 under RF. For marital status, recall was 0.69 for not-married and 0.61 for married employees under LR, and 0.79 against 0.56 under RF. The two measures can disagree: the demographic parity difference of RF for gender is almost zero (0.0017) while its equalized odds difference is 0.1169, so a parity check alone would have passed this model.

D. Bias Mitigation

Both methods were applied to age group. Fig. 3 shows recall by age group before and after mitigation, and Table IV gives the two fairness differences.

Only one of four mitigations raised recall for employees 45 and above

Recall by age group for each model and mitigation (LR: Logistic Regression, RF: Random Forest), test set



test set, $n = 294$ (47 employees who left; 9 of them aged 45 and above), values from the notebook outputs

Fig. 3. Recall by age group for each model and mitigation. The recall gap equals the equalized odds difference in every configuration, because the true-positive-rate gap was the larger component.

Table IV. Demographic parity difference (DPD) and equalized odds difference (EOD) for age group, before and after mitigation.

Configuration	DPD	EOD
LR	0.1973	0.4286
LR + ThresholdOptimizer	0.0671	0.3007
LR + ExponentiatedGradient	0.1019	0.1111
RF	0.2654	0.4902
RF + ThresholdOptimizer	0.1409	0.2941
RF + ExponentiatedGradient	0.0882	0.1765

All four mitigated models narrowed the recall gap and lowered the equalized odds difference. However, three of the four did so by lowering recall for the 45-and-above group. ThresholdOptimizer reduced it from 0.33 to 0.11 for LR and to zero for RF. ExponentiatedGradient reduced it to zero for RF, where recall for the other groups also fell sharply



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

(0.05 for 30 to 45 and 0.18 for under 30). Only LR + ExponentiatedGradient lowered the equalized odds difference from 0.43 to 0.11 while raising recall for the 45-and-above group from 0.33 to 0.56; recall for the other two groups fell from 0.76 to 0.67 (30 to 45) and from 0.71 to 0.65 (under 30).

E. Discussion

Three of the four mitigations show levelling down [11]: the metric improved because the better-served groups lost recall, or because every group did, not because the 45-and-above group was served better. A tool that met the fairness criterion in this way would flag almost none of the at-risk older employees, which defeats the purpose of the audit. This is why subgroup recall is reported next to every fairness metric. LR + ExponentiatedGradient is the one configuration in which the gap narrowed and the protected group gained. The gain corresponds to two more of the nine employees who left being flagged (from 3 of 9 to 5 of 9), and the other groups lost some recall. With so few cases it should be read as a lead to confirm with repeated splits, not as an established result. Why the two methods behaved differently was not tested here. One possible factor is that ThresholdOptimizer was run with its default objective, which maximizes accuracy; with 16 percent positives, an accuracy objective favors negative predictions, and an objective that values recall, such as balanced accuracy, may behave differently. Three practical points follow. First, report recall for each group alongside any summary fairness metric. Second, compare more than one mitigation method, since results differed by method and by model in this study. Third, ThresholdOptimizer needs the sensitive attribute at prediction time, which may be unacceptable for a protected attribute such as age in hiring, whereas the ExponentiatedGradient model was used for prediction without it.

VI. LIMITATIONS AND THREATS TO VALIDITY

1. **Proxy outcome.** Attrition stands in for role fit, but employees leave for many reasons unrelated to fit, and no psychometric scores are used. A public Big Five dataset was examined, but it contains no demographic attributes and so could not support a fairness audit.
2. **Fictional dataset.** The data are fictional [5], [16]. The findings illustrate a method and should not be read as evidence about a real employer.
3. **Single split and small subgroups.** One stratified split (seed 42) was used, with no repeated runs or confidence intervals. The 45-and-above test subgroup has nine employees who left, so a difference of one or two cases changes recall by 11 to 22 points.
4. **Threshold chosen on the test split.** The threshold of 0.3 was chosen after inspecting test-set results, without a validation set, so the audited Random Forest may look better than it would on new data.
5. **Mitigation comparison is not like for like.** The mitigated Random Forest models were fitted on the original training data, whereas the baseline used SMOTE and a 0.3 threshold. ThresholdOptimizer was run with its default accuracy objective, which may favor negative predictions on imbalanced data; other objectives were not tested.
6. **Scope of mitigation.** Mitigation was applied to age group only. Gender and marital status were not re-audited after mitigation, and the overall accuracy, precision and recall of the mitigated models were not recorded.
7. **Sensitive attribute at prediction time.** ThresholdOptimizer needs age at prediction time; the legal and ethical acceptability of this was not assessed.
8. **Grouping choices.** The age bands and the married versus not-married indicator are analyst choices, and other groupings could change the numbers. The associations reported are not causes.

VII. CONCLUSION

This paper audited two attrition-prediction models, used as a proxy for behavioral role-fit assessment, across gender, age group and marital status, and compared two mitigation methods. Age group showed the largest disparity: employees aged 45 and above who left were flagged far less often (recall 0.33) than younger employees (0.71 to 0.82), and the equalized odds difference reached 0.43 to 0.49. Every mitigation narrowed the gap, but three of the four did so by lowering recall for the 45-and-above group, in two cases to zero. Only Logistic Regression with ExponentiatedGradient narrowed the gap (0.43 to 0.11) and raised recall for that group (0.33 to 0.56).

The main contribution is a worked example showing that a fairness metric can improve while the group it is meant to protect is served worse, so mitigation should be judged on the metric and on the protected group's recall together. Because the results rest on one split with few positive cases in the key subgroup, they are indicative and need



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

confirmation. Future work will repeat the experiments over repeated stratified splits with confidence intervals, tune the decision threshold on a validation set, test ThresholdOptimizer with a recall-oriented objective, add pre-processing mitigation and other constraints such as equal opportunity, re-audit all attributes after mitigation, and evaluate on data that pair real psychometric assessment scores with demographic fields and role-fit outcomes.

DATA AVAILABILITY AND CODE REPOSITORY

The IBM HR Analytics dataset is publicly available on Kaggle [16]. The notebook used for exploratory analysis, preprocessing, modeling, auditing and mitigation is maintained in a GitHub repository: [repository URL to be added by the authors].

VIII. ACKNOWLEDGMENT

Purushothamman Elango Pillai thanks the faculty of the Department of Data Science at SIES College of Arts, Science, and Commerce, Mumbai, for academic guidance.

REFERENCES

- [1] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, Oct. 10, 2018.
- [2] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAT*), 2020, arXiv:1906.09208.
- [3] S. Yang and M. T. Islam, "IBM employee attrition analysis," arXiv preprint arXiv:2012.01286, 2021.
- [4] D. Liu, Z. Shafi, W. Fleisher, T. Eliassi-Rad, and S. Alfeld, "RAWLSNET: Altering Bayesian networks to encode Rawlsian fair equality of opportunity," arXiv preprint arXiv:2104.03909, 2021.
- [5] A. Fabris, S. Messina, G. Silvello, and G. A. Susto, "Algorithmic fairness datasets: The story so far," Data Mining and Knowledge Discovery, 2022, arXiv:2202.01711.
- [6] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products," in Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES), 2019, pp. 429–435.
- [7] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," arXiv preprint arXiv:1908.09635, 2019.
- [8] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning. fairmlbook.org, 2019.
- [9] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems (NIPS), 2016, pp. 3315–3323.
- [10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in Proc. 3rd Innovations in Theoretical Computer Science Conf., 2012, pp. 214–226.
- [11] B. Mittelstadt, S. Wachter, and C. Russell, "The unfairness of fair machine learning: Levelling down and strict egalitarianism by default," arXiv preprint arXiv:2302.02404, 2023.
- [12] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, white paper, 2020. [Online]. Available: <https://aka.ms/fairlearn-whitepaper>
- [13] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, "Fairlearn: Assessing and improving fairness of AI systems," J. Mach. Learn. Res., vol. 24, no. 257, pp. 1–8, 2023.
- [14] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in Proc. Int. Conf. Machine Learning (ICML), 2018, pp. 60–69.
- [15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
- [16] P. Subhash, "IBM HR Analytics Employee Attrition & Performance," Kaggle dataset, 2017. [Online]. Available: <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>
- [17] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
- [18] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," J. Mach. Learn. Res., vol. 18, no. 17, pp. 1–5, 2017.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com