World Journal of Management Science
Print ISSN: 2959-9628
Online ISSN: 2959-9636
DOI: https://doi.org/10.61784/wms4121

FLEX-TALENT: AN EXPLAINABLE AND FAIR MACHINE LEARNING FRAMEWORK FOR HIGH-STAKES LEADERSHIP-PIPELINE EVALUATION

ZiMeng Wang[1*], ZiJian Shen[2]
[1]*Brandeis University, Waltham 02453, MA, USA.*
[2]*Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[*]*Corresponding Author: ZiMeng Wang*

Abstract: Machine learning can prioritize candidates for leadership pipelines, but historical promotion records encode organizational policy rather than objective executive potential. This distinction is critical in high-stakes talent evaluation, where a technically accurate model may reproduce label bias, obscure unequal error burdens, or invite automation bias. This paper presents FLEX-Talent, a construct-aware, budget-constrained framework that combines feature governance, training-only fairness reweighting, predeclared validation gates, group calibration audits, intersectional testing, and explanation-stability checks. The framework is evaluated on a real public HR promotion dataset containing 54,808 employees and 4,668 historical promotions. Gender is never used at inference; it is retained only for auditing and reweighting. Five candidate models are compared under a fixed 10% shortlist budget. The validation rule selected a reweighted CatBoost model, which achieved 0.9134 ROC-AUC and 0.6116 average precision on an untouched test set. However, its absolute gender equal-opportunity difference increased from 0.0138 on validation to 0.0578 on test, slightly exceeding the predeclared 0.05 gate; a 95% bootstrap interval of [0.0060, 0.1497] reveals substantial fairness uncertainty. Rather than treating this as a hidden failure, FLEX-Talent makes fairness generalization a first-class release decision. Global and local SHAP evidence is accompanied by bootstrap stability analysis and explicit non-causal warnings. The resulting contribution is not an automated promotion system, but a reproducible governance architecture for human review, appeal, and post-deployment monitoring.
Keywords: Algorithmic HRM; Executive talent evaluation; Explainable AI; Fairness; Promotion prediction; CatBoost; SHAP; Responsible AI

1 INTRODUCTION

Organizations increasingly use predictive analytics to identify employees for promotion, succession planning, and accelerated development. The attraction is understandable: a model can aggregate heterogeneous performance, tenure, training, and career-history variables consistently and at a scale that unaided committees cannot reproduce. Yet executive and leadership-pipeline decisions are not ordinary classification problems. They allocate scarce visibility, development resources, compensation, and future authority. Errors can compound over a career, while apparently neutral historical labels may reflect earlier managerial discretion, opportunity access, and organizational politics. Reviews of artificial intelligence in human resource management therefore emphasize the gap between technical promise and organizational reality [1-6].

The central measurement problem is that an observed promotion label is not a direct observation of executive talent. It is an outcome generated by a prior decision process. A model trained on this label learns a conditional approximation to historical promotion policy, including both useful signal and potentially undesirable regularities. Removing gender from the feature vector prevents direct dependence on gender, but it does not remove proxies, differential measurement error, or historical label bias. Conversely, enforcing a single parity statistic can degrade calibration or conceal other error asymmetries; several fairness criteria are mutually incompatible except under restrictive conditions [7-13].

This paper addresses these issues through FLEX-Talent, an explainable and fair machine learning framework for high-stakes leadership-pipeline evaluation. FLEX denotes five linked commitments: Feature and construct governance, Label-aware fairness, EXplanation evidence, and lifecycle control. The system is explicitly a shortlist prioritizer for structured human review, not an automated promotion decision maker. It uses protected attributes only in a separated audit channel, compares full and governed feature sets, optionally applies Kamiran-Calders reweighting during training, selects a model under a fixed operational budget using predeclared validation constraints, and tests both fairness point estimates and uncertainty on an untouched test set.

The empirical contribution is deliberately falsifiable. On 54,808 real employee records, the validation gate selected reweighted CatBoost because it had the highest validation average precision among models meeting the declared fairness and group-calibration constraints. On test, ranking performance remained strong, but equal-opportunity parity weakened enough to cross the release boundary. This validation-to-test discrepancy is not corrected by choosing a different model after observing test results. Instead, it motivates an uncertainty-aware deployment decision: fairness is a statistical property that must generalize, not a checkbox satisfied on a single split.

The paper makes four contributions. First, it introduces a construct gate that prevents historical promotion from being presented as ground-truth leadership potential. Second, it proposes a budget-aware validation gate that jointly controls ranking utility, opportunity parity, disparate impact, and group calibration before the test set is opened. Third, it adds explanation consensus: global SHAP, local evidence cards, and bootstrap ranking stability are reported together and explicitly interpreted as descriptive rather than causal. Fourth, it demonstrates an honest fairness-generalization audit in which the selected model's test violation and confidence interval become governance evidence rather than a result to suppress.

2 RELATED WORK

2.1 Algorithmic HRM and High-Stakes Talent Decisions

Algorithmic HRM research identifies both organizational opportunities and fundamental constraints. Tambe et al. argue that HR data are often sparse, strategic behavior changes after measurement, and causal inference is difficult [1]. Leicht-Deobald et al. connect algorithm-based HR decisions to personal integrity and the risk that efficiency-oriented systems narrow acceptable employee behavior [2]. Köchling and Wehner's systematic review documents discrimination concerns across recruitment and development, while Raghavan et al. show that vendor claims about debiasing hiring systems often lack transparent validation [3-4]. Work on automation and augmentation further shows that affected employees and third-party observers react not only to outcomes but also to the process, locus of control, and perceived legitimacy [5, 14]. These findings imply that predictive performance alone is an inadequate release criterion.

FLEX-Talent translates this socio-technical literature into model-development controls. It distinguishes the construct of interest from the available label; limits the model to review prioritization; records feature rationales; requires uncertainty-aware subgroup evaluation; and produces evidence that a committee can contest. This design aligns with the lifecycle orientation of the NIST AI Risk Management Framework and with dataset/model documentation practices [15-17].

2.2 Fair Machine Learning

Group fairness has multiple legitimate but non-equivalent definitions. Demographic parity compares selection rates, while equal opportunity compares true-positive rates conditional on the observed positive label [7]. Disparate-impact ratios provide an intuitive selection-rate diagnostic, but neither parity notion proves individual fairness or absence of causal discrimination. Chouldechova and Kleinberg et al. formalize incompatibilities among error balance, predictive parity, and calibration when base rates differ [8-9]. Pleiss et al. similarly analyze the tension between calibration and error-rate parity [13]. Consequently, FLEX-Talent treats fairness as a vector of monitored quantities rather than a single optimized number.

Mitigation can occur before, during, or after model training. Disparate-impact repair, reweighting, and reductions-based constrained optimization are influential approaches [10-12]. This study uses reweighting because it is model-agnostic, auditable, and separates the protected attribute from inference. The method equalizes the contribution of gender-label cells in the training objective without modifying labels. This choice is intentionally modest: it can change the empirical training distribution, but it cannot establish that the label is fair or that proxies have been removed. Toolkits such as AI Fairness 360 support broad metric and mitigation audits, but metrics still require a decision context and an accountable governance process [18-21].

2.3 Explainability and Evidentiary Limits

Post-hoc explanation methods can summarize influential features locally or globally, but high-stakes deployment requires caution [22-24]. Rudin argues that interpretable models should be preferred when feasible, whereas surveys show that the fidelity and semantics of explanations vary substantially [25-26]. Counterfactual explanations may offer actionable contrastive information, but an actionable change is not necessarily causally valid or normatively appropriate [27]. Moreover, post-hoc methods can be manipulated, and saliency-style explanations require sanity checks [28-29]. Social-science evidence indicates that people seek contrastive, selective, and audience-dependent explanations [30]. FLEX-Talent therefore avoids presenting a SHAP value as a reason why a person deserves promotion. It is evidence about the model's behavior under the fitted historical-policy approximation.

3 PROBLEM FORMULATION

3.1 Decision Scope and Construct Gate

Let X denote observed employee attributes available before a promotion cycle, A the protected audit attribute (gender in the case study), Y the historical promotion decision, and S=f(X) a score. The operational decision is not to predict an abstract executive-quality variable. It is to prioritize a fixed fraction b of employees for additional structured review. The deployed recommendation is D=1[S>=t_b], where t_b is selected on the validation distribution to approximately satisfy P(D=1)=b. In this study b=0.10. The construct gate records the stronger claim that is not justified: Y cannot be

treated as a direct, unbiased measure of latent executive talent. This boundary changes the appropriate use from automated selection to evidence-assisted review.

$$D_i = 1[S_i \geq t_b], \quad t_b = Q_{1-b}(S_{\text{validation}}) \tag{1}$$

The audit attribute A is never an inference input. It is stored in a separate evaluation channel and is used in the reweighting calculation on training data. This architecture guarantees direct counterfactual invariance to changing the gender field while holding model inputs fixed. It does not guarantee causal fairness because X may contain proxies or measurements shaped by unequal opportunities.

3.2 Training-Only Reweighting

For each training observation with protected group a and historical label y, FLEX-Talent optionally applies the Kamiran-Calders weight [11]

$$w(a, y) = \frac{P(A=a)P(Y=y)}{P(A=a,Y=y)}. \tag{2}$$

The weights are normalized to mean one and enter the learner's loss. Intuitively, overrepresented group-label cells contribute less and underrepresented cells contribute more. Because the transformation is limited to the training objective, deployment requires neither A nor a group-specific threshold. This avoids explicit disparate treatment at inference, while preserving a transparent audit trail of the intervention.

3.3 Validation Gate

Candidate models are assessed at the same shortlist budget and are eligible only when three validation conditions hold. Absolute equal-opportunity difference is |TPR_f-TPR_m|<=0.05; disparate-impact ratio is constrained to 0.80<=SR_f/SR_m<=1.25; and the absolute gender Brier-score gap is <=0.02. Among feasible candidates, the model with the highest average precision is selected. These thresholds are governance choices, not universal definitions of fairness. They are predeclared so the untouched test set cannot be used to tune either the model or the fairness rule.

$$m^* = \text{argmax}_m \text{AP}_m \text{ subject to } C_{\text{EOD}}, C_{\text{DIR}}, C_{\text{Brier}} \text{ on validation.} \tag{3}$$

Average precision is emphasized because only 8.52% of employees carry a positive historical promotion label; precision-recall analysis is more informative than ROC curves alone under severe class imbalance [31-32]. Brier score complements ranking metrics by penalizing poorly calibrated probabilities. For the selected model, the same threshold is then transferred to the test set. Bootstrap intervals quantify the sampling uncertainty of average precision, F1, absolute equal-opportunity difference, and disparate-impact ratio, see Figure 1.

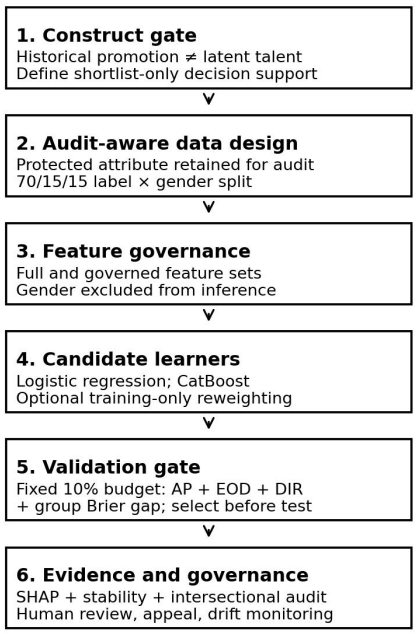

Figure 1 FLEX-Talent Workflow

Note: Protected attributes are isolated from inference and retained for training-time reweighting and audit only.

3.4 Explanation Consensus and Human Review

FLEX-Talent produces three levels of model evidence. Global mean absolute SHAP values summarize the fitted model's dominant associations. Local evidence cards list positive and negative contributions for representative true-positive, false-positive, and false-negative cases. Finally, 200 bootstrap resamples of the held-out explanation

sample evaluate whether the global feature ranking is stable. A stable explanation can still describe an undesirable policy, so stability is treated as reliability evidence rather than ethical validation.
The final decision remains with a structured committee. Reviewers receive the score band, evidence card, data-quality flags, and a warning that explanations are non-causal. They must document job-related evidence outside the model, can override recommendations with reasons, and must provide an appeal and correction route. Aggregate overrides and appeals become monitoring inputs rather than silent exceptions.

4 EXPERIMENTAL DESIGN

4.1 Dataset and Preprocessing

The experiment uses the public HR Analytics promotion dataset distributed for a machine-learning challenge. The verified file contains 54,808 unique employee identifiers and 14 columns, including 11 candidate predictors, gender, the target, and an identifier. There are 4,668 historical promotions (8.52%). The female subgroup contains 16,312 records and the male subgroup 38,496; historical promotion rates are 8.99% and 8.32%, respectively. Education is missing in the source data for 2,409 records and previous-year rating for 4,124 records. In the integrity-checked copy used here, categorical missing education values are represented by a missing token and numeric missing ratings are handled by the learner or median imputation, depending on the model. Employee identifiers are excluded from training.
The data are split 70/15/15 into training (38,365), validation (8,221), and test (8,222) sets, stratified jointly by Y and gender. The test set remains untouched until model and threshold selection are complete. This is a cross-sectional benchmark: no timestamp is available, so temporal drift cannot be estimated from the dataset and is listed as a deployment requirement rather than simulated, see Table 1.

Table 1 Dataset and Predeclared Protocol

Item	Value
Records / positives	54,808 / 4,668 (8.52%)
Protected audit groups	Female 16,312; Male 38,496
Split	38,365 / 8,221 / 8,222
Full features	11; excludes ID and gender
Governed features	8; also removes age, region, recruitment channel
Operational threshold	Top 10% validation shortlist
Random seed	42

4.2 Feature Governance and Candidate Models

The full feature set contains department, region, education, recruitment channel, number of trainings, age, previous-year rating, length of service, KPI-above-80% indicator, awards indicator, and average training score. The governed set excludes age, region, and recruitment channel because they may be weakly job-related or proxy-rich in a generic cross-organization setting. This is a sensitivity analysis, not a legal conclusion; a real organization would require role-specific job analysis and documented necessity for every feature.
Five candidates are evaluated: governed logistic regression; full CatBoost; full CatBoost with reweighting; governed CatBoost; and governed CatBoost with reweighting. Logistic regression uses one-hot encoding, median/mode imputation, standardization for numeric variables, and L2 regularization. CatBoost uses ordered categorical handling, depth six, learning rate 0.045, up to 400 iterations, early stopping on validation AUC, and a fixed random seed [33]. Hyperparameters are held constant across fairness variants to isolate the effect of reweighting and feature governance.

4.3 Metrics, Uncertainty, and Explanations

Ranking performance is reported with ROC-AUC and average precision. Thresholded performance at the transferred 10% budget includes precision, recall, F1, balanced accuracy, and Brier score. Gender audits include selection rates, demographic-parity difference, disparate-impact ratio, group TPR and FPR, equal-opportunity difference, precision gap, average-odds difference, and group Brier gap. An intersectional audit crosses gender with three age bands (<=30, 31-40, and >=41). Cells are reported with sample counts to prevent a small-group estimate from appearing as equally reliable as a large-group estimate.
For the selected model, 120 nonparametric bootstrap resamples of the test set produce percentile 95% intervals. Global SHAP uses 800 held-out cases. Explanation stability is measured on 200 bootstrap resamples of that explanation sample using top-five Jaccard overlap and Spearman rank correlation against the full-sample ranking. All values in the paper are generated by the supplied code; no table entry is manually invented.

5 RESULTS

5.1 Validation Selection

At the 10% validation budget, all five candidates satisfied the declared validation constraints. The full reweighted CatBoost model achieved the highest validation average precision (0.6329) and was therefore selected before the test set

was evaluated. Its validation absolute equal-opportunity difference was 0.0138, disparate-impact ratio 1.0177, and absolute gender Brier gap 0.0054. Figure 2 shows that the governed reweighted model approached the 0.05 opportunity-gap boundary, while the selected model occupied a favorable validation performance-parity region.

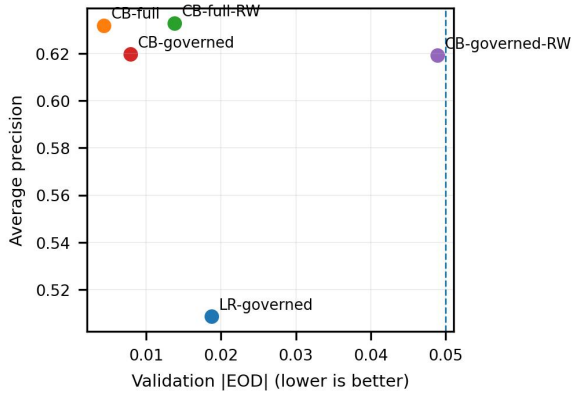

Figure 2 Validation Performance-Fairness Trade-Off under the Fixed 10% Shortlist Budget
Note: The dashed line is the predeclared |EOD| boundary.

5.2 Held-Out Predictive Performance

Table 2 reports the independent test results using each model's validation-derived threshold. The selected CB-full-RW model achieved ROC-AUC 0.9134, average precision 0.6116, precision 0.4638, recall 0.5214, and F1 0.4909. Its Brier score was 0.0480. Relative to governed logistic regression, average precision increased by 0.1107 and F1 by 0.0677. CatBoost's treatment of nonlinearities and categorical interactions therefore provided substantial ranking value in this benchmark.

The unweighted full CatBoost model has slightly higher test average precision (0.6120) and F1 (0.4970), but it cannot replace the selected model after test inspection. Doing so would leak the test set into model selection. The small difference also shows why the paper's contribution is not a claim that reweighting universally improves accuracy. Reweighting changed group behavior and the validation selection outcome; its effect must be evaluated empirically, see Figure 3.

Table 2 Held-Out Test Performance at Validation-Derived Thresholds

Model	AP	AUC	P	R	F1	Brier
LR-governed	0.5009	0.8694	0.4005	0.4486	0.4232	0.0573
CB-full	0.6120	0.9134	0.4668	0.5314	0.4970	0.0481
CB-full-RW	0.6116	0.9134	0.4638	0.5214	0.4909	0.0480
CB-governed	0.6008	0.9047	0.4716	0.5214	0.4953	0.0485
CB-governed-RW	0.6002	0.9045	0.4558	0.5229	0.4870	0.0487

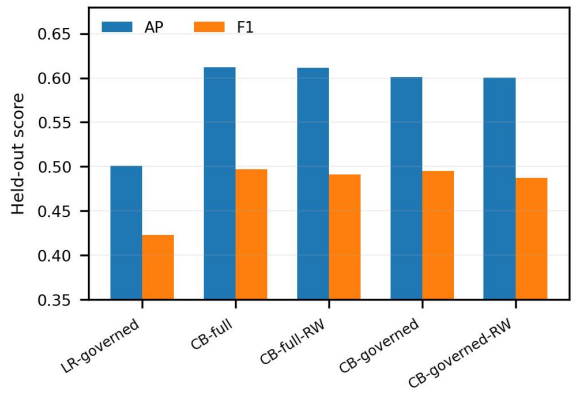

Figure 3 Held-Out Average Precision and F1
Note: CB-full-RW is the model selected strictly from validation data.

5.3 Fairness Generalization and Uncertainty

The selected model shortlisted 9.57% of test employees. Female and male selection rates were 10.75% and 9.07%, yielding a disparate-impact ratio of 1.1845, within the declared interval. Group Brier scores differed by only 0.0060. However, female TPR was 0.4818 and male TPR 0.5396, so the absolute equal-opportunity difference was 0.0578. This

exceeds the 0.05 validation gate by 0.0078. The failure is small as a point estimate but substantively important because the gate was declared before test evaluation.

Bootstrap uncertainty reinforces the need for caution. The 95% interval for absolute equal-opportunity difference is [0.0060, 0.1497], and the interval for disparate-impact ratio is [1.0316, 1.3260]. Thus, a single favorable validation split does not establish stable fairness. The appropriate action is not to optimize against the test set, but to delay or constrain deployment, collect more locally representative validation data, predefine a new evaluation cycle, and monitor rolling intervals. This is the paper's central empirical finding: fairness constraints themselves are subject to sampling and distribution shift, see Table 3.

Table 3 Selected Model: Generalization from Validation to Test

Metric	Validation	Test	Test 95% CI
Average precision	0.6329	0.6116	[0.5774, 0.6449]
F1	0.5227	0.4909	[0.4566, 0.5185]
|Equal-opportunity diff.|	0.0138	0.0578	[0.0060, 0.1497]
Disparate-impact ratio	1.0177	1.1845	[1.0316, 1.3260]
Absolute Brier gap	0.0054	0.0060	Not bootstrapped

For comparison, unweighted full CatBoost happened to have a lower test absolute opportunity gap (0.0392) but a higher disparate-impact ratio (1.2227). Governed models did not dominate: removing age, region, and recruitment channel reduced average precision by approximately 0.011-0.012 and produced larger test opportunity gaps. This result illustrates two cautions. First, proxy removal by intuition does not guarantee group parity. Second, a model that looks fairer on one metric may be less favorable on another. Feature governance must be connected to job relevance and repeated empirical audit rather than assumed to solve fairness.

5.4 Intersectional Audit

Gender-only averages mask heterogeneous errors. Among female employees aged 31-40, TPR was 0.4344 and precision 0.3841, whereas female employees aged 41 or above had TPR 0.6296 and precision 0.4595. Male precision ranged from 0.4532 in ages 31-40 to 0.5658 at age 41 or above. Some cells contain relatively few historical positives; therefore these values are signals for review, not stable rankings of group treatment. A production system should establish minimum cell sizes, hierarchical uncertainty estimates, and an escalation rule when intersectional intervals are too wide, see Table 4.

Table 4 Intersectional Test Audit by Gender and Age Band

Group	n	Base	Select	TPR	Prec.
f/31-40	1200	0.102	0.115	0.434	0.384
f/<=30	787	0.090	0.112	0.507	0.409
f/>=41	460	0.059	0.080	0.630	0.459
m/31-40	2727	0.085	0.102	0.545	0.453
m/<=30	1926	0.090	0.088	0.517	0.529
m/>=41	1122	0.067	0.068	0.573	0.566

5.5 Explanation Evidence

Global SHAP importance identifies KPI-above-80% as the strongest model association (mean absolute SHAP 1.4033), followed by average training score (0.9636), department (0.4510), and previous-year rating (0.3431). Region and age remain lower-ranked but nonzero. The top-five feature set was identical across all 200 explanation bootstrap resamples, and mean Spearman rank correlation with the full-sample ranking was 0.9997. This stability suggests that the reported ranking is not an artifact of a particular 800-case explanation sample, see Figure 4.

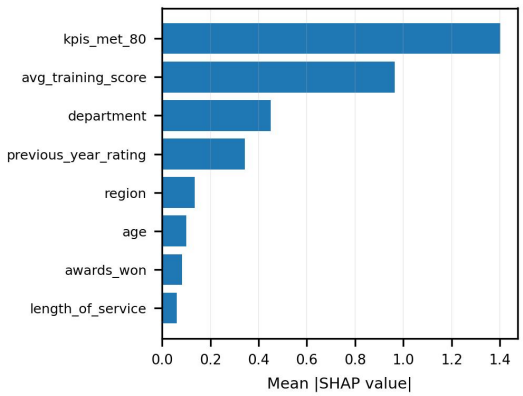

Figure 4 Global Mean Absolute SHAP Importance for the Selected Model

Note: Values describe model dependence, not causal merit.

Local evidence cards expose qualitatively different failure modes. A representative true positive received a near-one score driven by department and an exceptional training score, despite not meeting the KPI indicator. A false positive was strongly influenced by an award, high training score, and prior rating. A false negative just below the transferred threshold had a strong positive KPI contribution but weaker region, prior-rating, and award contributions. These cases are useful because reviewers can ask whether the data are accurate and job-related. They must not infer that changing a listed feature would causally change leadership ability or guarantee promotion.

6 DISCUSSION

6.1 What the Experiment Establishes

The experiment establishes that a real, imbalanced promotion dataset supports useful shortlist ranking and that fairness-aware model selection can be made reproducible. It also establishes that validation fairness may fail to generalize even when predictive ranking remains stable. The latter is easy to miss when papers report only a single test split or select a fairness intervention after inspecting test metrics. FLEX-Talent treats the test result as a release audit, so an out-of-bound statistic is evidence about deployment risk rather than an invitation to tune on the holdout.

The result does not establish that the selected model is fair in a legal, causal, or moral sense. Equal opportunity conditions on a historical label that may itself be biased. The direct gender-invariance rate is 1.0 by construction because gender is absent from the feature matrix, but proxy pathways can remain. Moreover, female test precision is 0.4030 versus 0.4943 for male employees, a gap of -0.0912. A committee that interprets all shortlisted cases identically could therefore impose different burdens of false review across groups even though selection-rate parity appears acceptable.

6.2 Deployment Protocol

A defensible deployment would begin with role-specific content validation. Each predictor should be linked to competencies required for the target leadership role, and variables with weak necessity should be removed even when they improve benchmark accuracy. The label should be expanded beyond a binary past decision to include structured assessment evidence, opportunity-adjusted performance, and longitudinal outcomes. Documentation should identify who can access scores, how long data are retained, and which decisions the model is prohibited from making.

During operation, the model should create a review queue rather than a final rank order. Human reviewers should use a standardized rubric, view uncertainty and data-quality warnings, and record independent evidence before seeing the model explanation where feasible to reduce anchoring. Employees should be able to inspect relevant source data, correct errors, and appeal. Monitoring should track subgroup and intersectional selection, TPR where outcomes become available, precision, calibration, overrides, appeals, and missingness. Release criteria should use confidence bounds and temporal windows, not point estimates alone. Material drift or a fairness-bound crossing should pause automated prioritization until a new predeclared validation cycle is completed.

6.3 Why Explanations Are Not a Fairness Certificate

Stable SHAP rankings improve reproducibility, but they do not prove that the model uses legitimate reasons. A historically biased policy can be explained consistently. Likewise, a local contribution is conditional on the fitted model and background distribution; it is not a causal estimate and may change under correlated-feature interventions. FLEX-Talent therefore presents explanations alongside feature-governance records, subgroup metrics, failure cases, and warnings. This evidence bundle is more useful than a single colorful explanation because it supports contestability: a reviewer can challenge the input, the relationship to the job, the threshold, or the population-level consequences.

6.4 Governance Artifacts and Accountability Allocation

Technical reproducibility is part of governance because it allows an independent reviewer to reconstruct how a candidate entered the queue. The accompanying package stores the dataset hash, split sizes, feature sets, model thresholds, validation and test metrics, bootstrap intervals, row-level test predictions, explanation samples, and the serialized selected model. A verification ledger records the bibliographic identity and publication year of every cited work. These artifacts separate claims generated by code from interpretive claims made by the author and reduce the risk that a later presentation silently changes the decision rule.

Accountability must also be assigned across roles. Data owners are responsible for provenance and correction; domain experts for construct and job relevance; model developers for reproducibility, testing, and limitation statements; the decision committee for independent evidence and documented overrides; and organizational leadership for appeal, monitoring, and authority to stop use. A model card can summarize performance and intended use, while a dataset datasheet can preserve collection and missingness context [15-16]. Neither document substitutes for an empowered owner who can halt the system when release conditions fail.

7 LIMITATIONS AND THREATS TO VALIDITY

The dataset is a public benchmark rather than an audited executive-succession dataset. It contains employees across departments and does not identify executive-level roles, organization, jurisdiction, compensation impact, or time. Accordingly, the case study supports leadership-pipeline shortlist methodology but cannot demonstrate validity for chief-executive selection. Historical promotion is an imperfect and potentially biased proxy, and no causal conclusions are drawn.

Only binary gender categories are present in the source data; the experiment cannot audit nonbinary identities or other protected dimensions such as race, disability, or socioeconomic background. Age is available and used for intersectional audit, but the public file does not provide sufficient context to determine whether age is job-related. The train-validation-test split is random because timestamps are absent, so performance and fairness under temporal drift may be worse than reported. Hyperparameter exploration is intentionally limited, and a different model family or constraint method could produce a different frontier.

Fairness statistics rely on the observed label. Equal opportunity can therefore reproduce inequity if access to prior opportunity influenced promotion. Bootstrap intervals quantify sampling variability conditional on this dataset, not uncertainty from label construction, omitted variables, organizational change, or strategic behavior. SHAP stability is measured for feature ranking, not explanation fidelity under all perturbations. Finally, the proposed human-review controls require organizational commitment; human involvement alone does not prevent automation bias or discriminatory overrides.

8 CONCLUSION

FLEX-Talent reframes high-stakes talent modeling as a governed evidence process rather than a search for a single accurate classifier. Its construct gate prevents historical promotion from being mislabeled as innate executive potential; its validation gate fixes the decision budget and fairness rules before test evaluation; and its explanation consensus combines global, local, and stability evidence with non-causal interpretation. The real-data experiment produced strong ranking performance but also a small test-time equal-opportunity violation with a wide confidence interval. That result is the practical lesson: fairness must generalize and remain monitored. The supplied code, predictions, metrics, model artifact, and verification ledger enable exact reproduction and critical examination. In deployment, the model should prioritize structured human review, never replace accountable judgment, and be paused when evidence falls outside predeclared bounds.

DATA AND CODE AVAILABILITY

The accompanying reproducibility package includes an integrity-checking dataset downloader, complete experiment code, pinned package versions, row-level held-out predictions, all reported machine-readable results, figures, the selected CatBoost model, and a reference-verification ledger. The dataset SHA-256 hash used in the experiment is recorded in the package. Redistribution and operational use remain subject to the dataset source's terms and applicable employment law.

REPRODUCIBILITY CHECKLIST

A clean reproduction should create an isolated environment, install the pinned dependencies, run the integrity-checking downloader, and execute the experiment once with seed 42. The reported model comparison is read directly from paper_main_results.csv; the selected-model metrics and intervals are read from selected_model_summary.json; and intersectional and explanation results have separate machine-readable files. The serialized model and held-out predictions permit independent recomputation of thresholded metrics without retraining. A reproduction is considered concordant only when the dataset hash, split sizes, selected model, validation threshold, and rounded paper values agree.

For a new organization or promotion cycle, exact numerical agreement is neither expected nor desirable. The organization should create a new immutable dataset snapshot, preserve a temporal holdout where possible, re-run construct and feature review, predeclare the shortlist budget and fairness bounds, and record the code revision and approving owners. The resulting decision record should include failed gates as well as passed gates. This distinction separates computational reproducibility of the present study from external validity in a different workforce and prevents a benchmark result from being reused as a deployment certificate.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Tambe P, Cappelli P, Yakubovich V. Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 2019, 61(4): 15-42. DOI: 10.1177/0008125619867910.

[2] Leicht-Deobald U. The challenges of algorithm-based HR decision-making for personal integrity. Journal of Business Ethics, 2019, 160(2): 377-392. DOI: 10.1007/s10551-019-04204-w.
[3] Köchling A, Wehner M C. Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 2020, 13: 795-848. DOI: 10.1007/s40685-020-00134-w.
[4] Raghavan M, Barocas S, Kleinberg J, et al. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*), 2020: 469-481. DOI: 10.1145/3351095.3372828.
[5] Langer M, Landers R N. The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. Computers in Human Behavior, 2021, 123: Art. 106878. DOI: 10.1016/j.chb.2021.106878.
[6] Rodgers W, Murray J M, Stefanidis A, et al. An artificial intelligence algorithmic approach to ethical decision-making in human resource management processes. Human Resource Management Review, 2023, 33(1): Art. 100925. DOI: 10.1016/j.hrmr.2022.100925.
[7] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems 29, 2016.
[8] Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 2017, 5(2): 153-163. DOI: 10.1089/big.2016.0047.
[9] Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. In: Proceedings of the Innovations in Theoretical Computer Science Conference (ITCS), LIPIcs, 2017, 67: Art. 43. DOI: 10.4230/LIPIcs.ITCS.2017.43.
[10] Feldman M. Certifying and removing disparate impact. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2015: 259-268. DOI: 10.1145/2783258.2783311.
[11] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 2012, 33: 1-33. DOI: 10.1007/s10115-011-0463-8.
[12] Agarwal A, Beygelzimer A, Dudík M, et al. A reductions approach to fair classification. In: Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2018, 80: 60-69.
[13] Pleiss G, Raghavan M, Wu F, et al. On fairness and calibration. In: Advances in Neural Information Processing Systems 30, 2017: 5684-5693.
[14] Binns R. It's reducing a human being to a percentage: Perceptions of justice in algorithmic decisions. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), 2018: Art. 377. DOI: 10.1145/3173574.3173951.
[15] Mitchell M. Model cards for model reporting. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*), 2019: 220-229. DOI: 10.1145/3287560.3287596.
[16] Gebru T. Datasheets for datasets. Communications of the ACM, 2021, 64(12): 86-92. DOI: 10.1145/3458723.
[17] Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology, 2023. DOI: 10.6028/NIST.AI.100-1.
[18] Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of the ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO), 2021: Art. 17. DOI: 10.1145/3465416.3483305.
[19] Selbst A D. Fairness and abstraction in sociotechnical systems. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*), 2019: 59-68. DOI: 10.1145/3287560.3287598.
[20] Jacobs A Z, Wallach H. Measurement and fairness. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2021: 375-385. DOI: 10.1145/3442188.3445901.
[21] Bellamy R K E. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 2019, 63(4/5): 4:1-4:15. DOI: 10.1147/JRD.2019.2942287.
[22] Ribeiro M T, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016: 1135-1144. DOI: 10.1145/2939672.2939778.
[23] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, 2017: 4765-4774.
[24] Lundberg S M. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020, 2: 56-67. DOI: 10.1038/s42256-019-0138-9.
[25] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1: 206-215. DOI: 10.1038/s42256-019-0048-x.
[26] Guidotti R,. A survey of methods for explaining black box models. ACM Computing Surveys, 2018, 51(5): Art. 93. DOI: 10.1145/3236009.
[27] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 2018, 31(2): 841-887.
[28] Slack D. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2020: 180-186. DOI: 10.1145/3375627.3375830.
[29] Adebayo J. Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems 31, 2018: 9505-9515.

[30] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2019, 267: 1-38. DOI: 10.1016/j.artint.2018.07.007.
[31] He H, Garcia E A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284. DOI: 10.1109/TKDE.2008.239.
[32] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 2015, 10(3): Art. e0118432. DOI: 10.1371/journal.pone.0118432.
[33] Prokhorenkova L. CatBoost: Unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems 31, 2018: 6638-6648.