The American Journal of Applied Sciences
ISSN 2689-0992 | Open Access

TYPE Original Research
PAGE NO. 56-65

Check for updates

Data-Driven Retention Targeting: A Holistic Analytics Framework Spanning Prediction, Causality, and Fairness

Nidhi Singh

Senior Data Analyst, State of Alabama, AL USA

OPEN ACCESS

SUBMITED 15 July 2024
ACCEPTED 17 August 2024
PUBLISHED 24 September 2024
VOLUME Vol.06 Issue 09 2024

CITATION
Singh, N. (2024). Data-Driven Retention Targeting: A Holistic Analytics Framework Spanning Prediction, Causality, and Fairness. The American Journal of Applied Sciences, 6(09), 56–65. Retrieved from https://theamericanjournals.com/index.php/tajas/article/view/8015

Abstract: Attrition entails significant costs of hiring, lost productivity, and lost know-how, which drive ML research on employee attrition prediction. Nevertheless, most existing work offers just one discriminatory statistic on one IBM HR Analytics attrition prediction synthetic split, without providing much guidance, and hardly addresses cost, interpretability, dynamics over time, robustness to new data, and fairness altogether. In contrast, this paper proposes a holistic decision-oriented framework and a new targeting policy contributing to attrition analysis in the following six dimensions: (i) cost-sensitive stacked ensemble (LightGBM, CatBoost, logistic regression) with repeated cross-validation, confidence intervals, and expected net savings metric rooted in retention economics; calibration proves indispensable for any cost-sensitive applications; (ii) post hoc explainability based on SHAP(SHapley Additive exPlanations) explanations together with DiCE (Diverse Counterfactual Explanations)counterfactual recourse; (iii) survival analysis (Kaplan-Meier estimator, Cox proportional hazards model) applied to a time-to-event dataset of turnover as another base-classification target; (iv) uplift modeling using three kinds of learners (S-, T-, and X-); (v) Fairness-Aware Cost-Sensitive Retention Targeting policy, FACS-RT, integrating uplift, cost, and fairness optimization in one algorithm and constructing value-fairness Pareto frontier; and (vi) leave-one-department-out resampling and auditing with respect to group fairness criterion. For the IBM dataset (n = 1,470), our approach yields an average AUC of 0.83 with 95%

confidence interval (0.79–0.89) with cross-validation, statistically equivalent to a strong logistic regression baseline (paired-bootstrap test p = 0.37), and isotonic calibration brings ECE down to 0.04. For the turnover dataset (n = 1,129), our method achieves AUC 0.72 and Cox concordance 0.66. There is a partial agreement between risk- and uplift-based ranking orders of 27%. FACS-RT retains 83% of expected maximum value while decreasing gender disparity of selection rates by 92% (0.053 → 0.004).

Keywords: Explainable AI; SHAP; Counterfactual explanation; Survival data analysis; Uplift modeling; Cost-sensitive learning; Algorithmic fairness

I. INTRODUCTION

Voluntary employee attrition is among the most costly and disruptive events an organization faces. Beyond the direct expense of recruiting and onboarding a replacement—commonly estimated at one-half to two times the departing employee's annual salary—organizations absorb productivity loss during vacancy, the erosion of institutional knowledge, and the destabilization of teams. Consequently, data-driven attrition analytics has become a central topic in human resource (HR) analytics and applied machine learning.

The most common methodology in the field trains a supervised classifier on the IBM HR Analytics: Employee Attrition and Performance dataset [22] and measures only a single discriminative metric such as the area under the receiver-operating-characteristic curve (AUC). Although this practice works well as a benchmark, it has several known problems. First, AUC does not account for intervention economics, so an excellent ranking can still produce losses for a firm, because intervening is costly and may target employees who would have stayed without any action. Second, a raw risk score alone leaves managers unable to make informed decisions about how to act once a risk is identified. Third, attrition is treated as a binary problem rather than as a time-to-event one, which misses an important dimension of the analysis. Fourth, the lack of transfer evaluation makes results overoptimistic about how well models generalize across different organizations. Fifth, algorithmic fairness is rarely studied, even though such algorithms help decide the fate of individual employees.

This paper argues for moving attrition analytics forward from accuracy to decision support. We combine six capabilities that are usually considered separately and demonstrate their integration on three publicly available datasets. Our specific contributions are as follows. First, we design a cost-sensitive stacked ensemble [17], [18], evaluate it with repeated stratified cross-validation and 95% confidence intervals, test the ensemble-versus-baseline gap with a paired bootstrap, and show that probability calibration is a prerequisite for cost-sensitive targeting. Second, we provide dual-mode explainability by combining SHAP [6] global and local attributions with DiCE [7] counterfactuals, turning opaque scores into auditable drivers and actionable recourse. Third, we add a survival-analysis track using the Kaplan–Meier estimator [8] and the Cox proportional-hazards model [9] on a structurally different turnover dataset that we also use as a second classification base, giving genuinely cross-dataset evidence rather than single-dataset results. Fourth, we estimate uplift with three meta-learners—the S-, T-, and X-learner [10]—report their rank agreement, and quantify the divergence between risk-based and uplift-based targeting. Fifth, as our principal methodological contribution, we introduce Fairness-Aware Cost-Sensitive Retention Targeting (FACS-RT), a tunable policy that unifies uplift, intervention economics, and a group-fairness constraint [12] into a single decision rule, with risk-first and value-first targeting as special cases, and we characterize its value–fairness Pareto frontier. Sixth, we assess cross-domain generalization via a leave-one-department-out protocol and conduct a group-fairness audit across gender and age bands.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the datasets. Section 4 presents the methodology. Section 5 reports experimental results. Section 6 discusses implications, Section 7 states limitations and threats to validity, and Section 8 concludes.

II. LITERATURE REVIEW

Supervised classifiers—logistic regression, decision trees, random forests, and gradient-boosted trees—dominate the attrition-prediction literature. Tree ensembles such as random forests [1] and gradient boosting machines including XGBoost [2], LightGBM [3],

and CatBoost [4] consistently rank among the strongest tabular learners and are the typical choice for the IBM dataset. Because attrition is class-imbalanced, studies employ class weighting or synthetic oversampling (SMOTE) [5]. A recurring limitation is single-split evaluation on one synthetic dataset, which limits external validity.

Because these models carry real implications for important personnel decisions, interpretation is key. SHAP [6] provides additively interpretable feature attributions with efficient tree-based estimates and is considered the state-of-the-art method for explaining attrition models. Counterfactual methods such as DiCE [7] complement this by suggesting the smallest changes that would alter a model's prediction, thereby enabling actionable remedies. While most human-resources research uses only one of these two approaches, the present work combines them.

Turnover is naturally a time-to-event process with right-censoring, since some employees are still present at the observation horizon. The Kaplan–Meier estimator [8] and the Cox proportional-hazards model [9] are classical tools that estimate survival curves and covariate hazards while respecting censoring—information that is discarded by binary classification. This survival framing answers "how soon" rather than only "whether," yet it remains comparatively underused in the applied attrition-ML literature.

Risk ranking identifies who is likely to leave, but the optimal allocation of a limited retention budget requires estimating who is most responsive to an intervention—the uplift, or conditional average treatment effect. Meta-learner approaches, namely the S-, T-, and X-learner [10], estimate heterogeneous effects from outcome models and are model-agnostic. In the absence of an experimental treatment column, intervention effects can be approximated by counterfactual reasoning over actionable features, an approach we adopt under a clearly stated assumption set. Beyond estimating uplift, allocating interventions is itself a constrained optimization problem, and integrating uplift with intervention economics and fairness constraints into a single targeting policy has received little attention in the HR setting.

Finally, when predicted probabilities drive expected-cost calculations, calibration—the agreement between predicted probabilities and observed frequencies—is as important as ranking quality; stacked and class-weighted models are frequently miscalibrated, and post-hoc methods such as isotonic regression and Platt scaling restore reliability. Separately, algorithmic decision-making in employment is subject to anti-discrimination scrutiny [11], with standard group-fairness criteria including demographic parity and equalized odds [12]. Auditing attrition models for disparate selection and error rates across protected groups is necessary before deployment but is frequently omitted, and fairness is almost never embedded directly in the retention-targeting decision.

RESEARCH GAP

To our knowledge, no prior study jointly addresses cost-sensitivity, calibration, dual-mode explainability, survival dynamics, multi-learner uplift, fairness-constrained targeting, cross-domain transfer, and statistical rigor on a common footing. We close this gap with an integrated framework, a novel fairness-aware targeting policy, and findings—such as the large divergence between risk and uplift rankings and the favorable value–fairness trade-off achievable by FACS-RT—that are obscured by prediction-only studies.

We use three publicly available datasets spanning two complementary modeling regimes. Table 1 summarizes their characteristics.

Table 1. Summary of the datasets used in this study.

Dataset	Records	Features	Role in this study
IBM HR Analytics	1,470	35	Primary: classification, SHAP/DiCE, uplift, fairness, cross-domain
Employee Turnover	1,129	16	Survival analysis (Kaplan–Meier, Cox PH); time-to-attrition
HR Employee Attrition	1,470	35	External-validation candidate (see note on identity below)

The IBM HR Analytics: Employee Attrition and Performance dataset [22] contains 1,470 employee records described by 35 features that span demographic factors (age, sex, marital status), compensation details (monthly salary, stock option level), job-related information (role, level, department, business travel), and survey-measured employee satisfaction and engagement. The binary target, Attrition, is positive in 16.1% of cases, reflecting a realistic class imbalance. Three departments are represented: Research & Development (961 employees), Sales (446 employees), and Human Resources (63 employees). It is generally accepted that IBM created this dataset synthetically, and the present work is therefore offered as a framework for analysis rather than as a study of an organic population.

The Employee Turnover dataset comprises 1,129 observations in which the dependent variable is the duration in months (stag) together with an event status (1 for turnover, 0 for censored). It also includes demographic, industrial, occupational, and psychometric variables such as extraversion, independence, self-control, anxiety, and novator. The event proportion is 50.6%, with a median duration of 24.3 months. The time-to-event nature of this dataset is precisely what classification methods ignore, which motivates the survival-analysis track.

III. METHODOLOGY

Our framework comprises seven stages: preprocessing and feature engineering; predictive modeling; cost-sensitive evaluation; explainability; survival analysis; uplift-based prescription; and cross-domain plus fairness assessment[14][15]. We detail each below.

A. Preprocessing and feature engineering

Constant and identifier columns (EmployeeCount, StandardHours, Over18, EmployeeNumber) are removed to prevent leakage and noise. We engineer six features motivated by HR theory: two tenure ratios (years in current role and years since last promotion, each normalized by company tenure), a log-transformed monthly income, a mean satisfaction index aggregating four survey items, a binary “stuck” flag (no promotion for ≥ 4 years with ≥ 5 years tenure), and an interaction term capturing overtime combined with low job satisfaction. Categorical variables are one-hot encoded with the first level dropped, and features are standardized. We use a stratified 80/20 train–test split.

B. Predictive models

We benchmark three class-weighted baselines—logistic regression, random forest, and XGBoost (with scale_pos_weight set to the negative/positive ratio)—against a focal **cost-sensitive stacked ensemble**. The ensemble stacks LightGBM, CatBoost, and logistic regression as base learners with a logistic-regression meta-learner over their predicted probabilities; all components use balanced class weighting. Stacking is chosen to combine the complementary inductive biases of gradient boosting and linear modeling while preserving calibrated probabilities for downstream cost computation [17].

C. Evaluation protocol, cost-sensitivity, and calibration

We evaluate discrimination with AUC, the precision–recall area, and F1, reporting **repeated stratified five-fold cross-validation** (five repeats) with means, standard deviations, and 95% percentile intervals, and we test the ensemble-versus-baseline AUC gap with a **2000-replicate paired bootstrap** on the test set. Because the downstream cost and value computations consume predicted probabilities, we additionally assess **calibration** via the expected calibration error and reliability diagrams, and apply isotonic regression when the raw model is miscalibrated. Discriminative metrics alone ignore the asymmetric economics of retention, so we define the expected net savings of contacting the top-k highest-risk employees as

$$ExpectedSavings(k) = p_r \cdot (\text{true leavers in top-}k) \cdot C_m - k \cdot C_i, \quad (1)$$

where C_m is the replacement cost of an unaddressed departure (US$20,000), C_i the per-employee intervention cost (US$5,000), and p_r the probability that an intervention retains a true leaver (0.5). Sweeping k traces the savings–budget frontier and identifies the value-maximizing operating point.

D. Explainability: SHAP and DiCE

For calculating the SHAP values, we rely on a tree-based explainer for the base learner LightGBM and evaluate the performance on global ranking by the mean of absolute contribution and local interpretation for the riskiest individual. In order to gain insight from our explanation, we create DiCE counterfactuals to

understand the feature changes required to change the prediction from "leave" to "stay."

E. Survival analysis

Regarding the analysis of the turnover data, we apply the kaplan-meier technique to estimate non-parametric survival, separately considering overall data and stratifying it by gender. We add a weak penalty to avoid unstable estimates in the cox regression modeling framework. Performance evaluation relies on the concordance index, equivalent of the auc in survival analysis.

F. Survival-versus-classification comparison

To compare paradigms fairly on identical data, we hold out 25% of the turnover dataset and train (a) a LightGBM classifier on the event label, excluding the duration variable, and (b) a Cox model on the same covariates. Both are scored by AUC against the event label and by the concordance index against observed durations, isolating the value of modeling time explicitly.

G. Uplift estimation with meta-learners

Because the data lack a randomized treatment column, we define a binary pseudo-treatment (absence of mandatory overtime, a recognized and actionable retention lever) and estimate the conditional effect of treatment on staying with three model-agnostic meta-learners: the S-learner (a single model with treatment as a feature), the T-learner (separate outcome models per arm), and the X-learner (imputed treatment effects combined through a propensity score). We report the Spearman rank agreement among the three uplift rankings as a stability check and use the T-learner uplift downstream. We stress that these estimates rest on a conditional-ignorability assumption that a field experiment would be required to verify; we therefore treat uplift as decision support, not causal proof[16].

H. Fairness-Aware Cost-Sensitive Retention Targeting (FACS-RT)

Our principal methodological contribution is a targeting policy that selects whom to contact under a fixed budget by jointly accounting for responsiveness, economics, and fairness. For employee i with estimated uplift u_i (reduction in leave probability under intervention), we define the expected net value of intervening as

$$v_i = u_i \cdot C_m - C_i, \qquad (2)$$

where C_m is the replacement cost and C_i the intervention cost. A value-first policy selects the top-k employees by v_i, and a risk-first policy is recovered by ranking on the raw risk score instead. FACS-RT augments value-first selection with a group-fairness constraint: employees are admitted in decreasing order of v_i, but an admission is rejected if it would push the absolute difference in per-group selection rate above a tolerance ε. Sweeping ε from strict to lose traces a **value–fairness Pareto frontier**, on which risk-first and value-first appear as limiting points. The policy thus exposes, rather than hides, the efficiency cost of fairness and lets HR choose an operating point explicitly. We evaluate policies by their model-estimated expected value (the ex-ante objective), their realized count of true leavers (ex-post validation against observed labels), and their gender selection disparity.

I. Cross-domain validation and fairness audit

Generalization is assessed with the leave-one-department-out protocol of Section 3.3, reporting external AUC and AUC-PR for each held-out department against the in-domain reference. For the standalone fairness audit, we compute per-group selection and recall rates and report demographic-parity and equalized-odds differences across gender, treating a gap above 0.05 as a mitigation trigger; age is examined via three bands.

IV. RESULTS

A, Experimental setup

All experiments run on a CPU environment with a fixed random seed (42) for reproducibility. Exploratory characteristics of the data are shown in Fig. 1: the IBM class imbalance, the age skew of leavers toward younger employees, and the turnover dataset's tenure-by-event distribution.

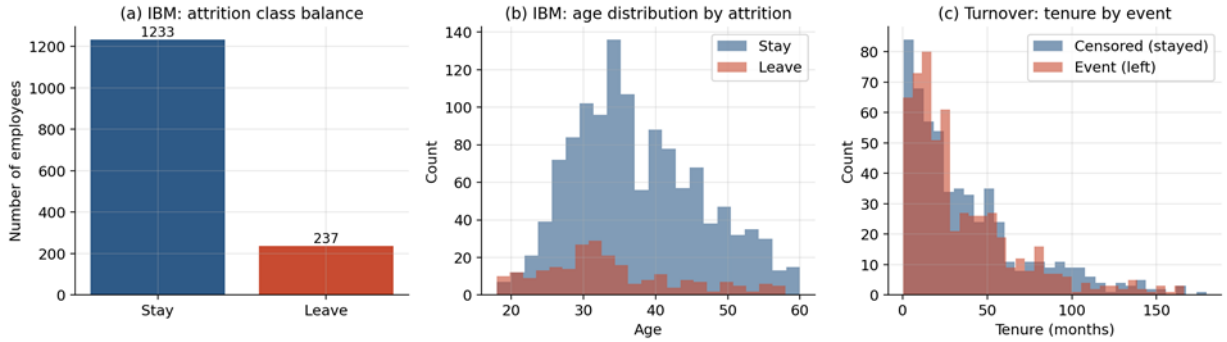

Fig. 1. Exploratory data analysis. (a) IBM attrition class balance; (b) age distribution by attrition; (c) turnover dataset tenure distribution by event status.

B. Predictive performance and statistical rigor

Table 2 reports discriminative performance under repeated stratified five-fold cross-validation (five repeats; mean, standard deviation, and 95% percentile interval), and Fig. 2 shows ROC and precision–recall curves on the held-out test split. The stacked ensemble attains the highest cross-validated AUC (0.832), narrowly ahead of logistic regression (0.830); the wide and overlapping intervals already suggest the gap is not material. A paired bootstrap on the test set confirms this: the ensemble–logistic AUC difference is −0.005 (95% CI [−0.015, +0.006], p = 0.37), i.e., statistically indistinguishable. We report this transparently—on this synthetic dataset a well-regularized linear model is a remarkably strong baseline—and show below that the ensemble's value lies in calibration and in serving as the substrate for cost-, uplift-, and fairness-aware decisions rather than in raw discrimination.

Table 2. Cross-validated discriminative performance on IBM (repeated 5-fold, 5 repeats): mean ± SD [95% interval]. Best mean per column in bold.

Model	AUC	AUC-PR	F1
Logistic Regression	0.830 ± 0.029	0.553 ± 0.062	0.499 ± 0.048
Random Forest	0.795 ± 0.039	0.490 ± 0.071	0.366 ± 0.071
XGBoost	0.797 ± 0.026	0.500 ± 0.057	0.456 ± 0.060
Stacked Ensemble	**0.832 ± 0.031**	**0.565 ± 0.064**	**0.508 ± 0.052**

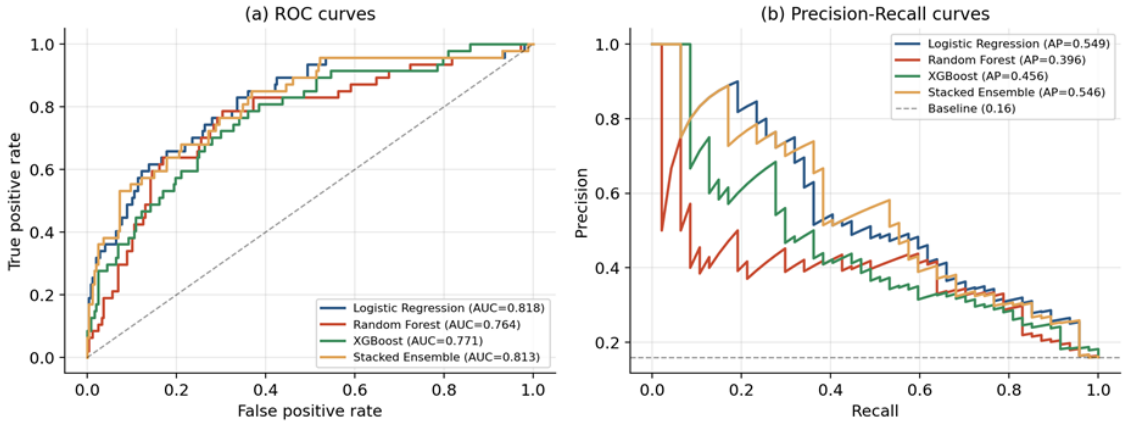

Fig. 2. Discriminative performance on the IBM test set: (a) ROC curves; (b) precision–recall curves with the prevalence baseline.

C. Cost-aware evaluation

Fig. 3 traces expected net savings as the retention budget (top-k) grows. At small, realistic budgets the ensemble and logistic regression yield the largest positive returns—peaking near +US$45,000 at k = 25 for the ensemble—after which savings decline and eventually turn negative as interventions are spent on employees who would have stayed. This frontier exposes an operating point that AUC alone cannot reveal and reframes model selection as a budget-aware decision.

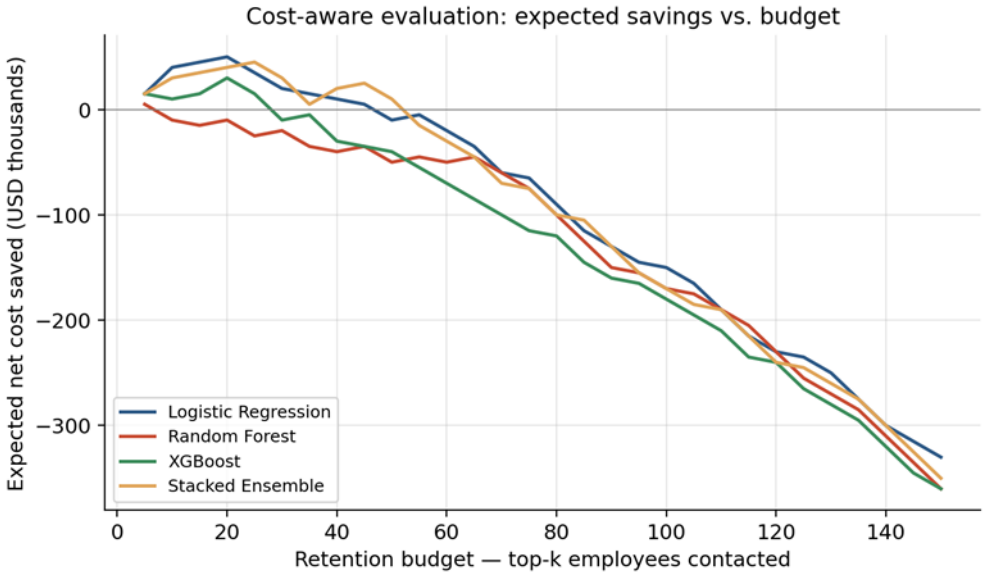

Fig. 3. Expected net cost saved (US$ thousands) versus retention budget. Positive returns concentrate at small budgets; over-targeting destroys value.

D. Probability calibration

Calibration is critical since calculations of both cost and value depend on predictions of probabilities. According to Figure 4, the raw stacked ensemble is highly optimistic with a curve much lower than the diagonal line and an ECE value of 0.22. Using isotonic regression decreases the ECE value to 0.04 and decreases the Brier score from 0.159 to 0.103, thus matching the probabilities to their respective frequency. This process is often overlooked in attrition analysis. However, failing to calibrate probabilities would shift the expected savings curve in Figure 3 to biased probabilities. Therefore, the ensemble is calibrated before any cost or value analysis is done.

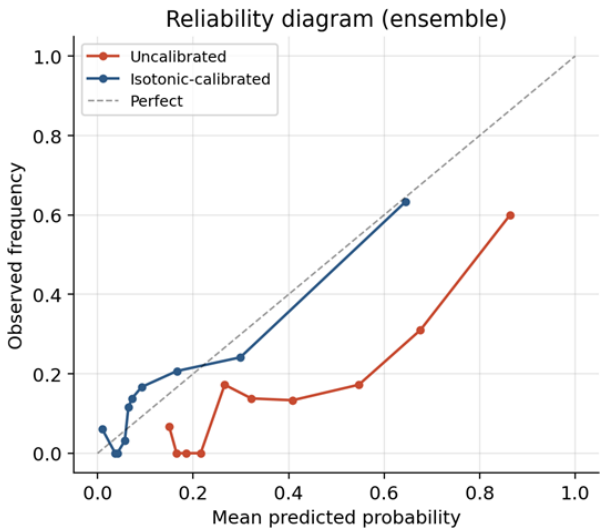

Fig. 4. Reliability diagram for the ensemble. The uncalibrated model is overconfident (ECE = 0.22); isotonic calibration restores reliability (ECE = 0.04).

E. Ablation study

Contribution of the elements of the framework is given in Table 3 below. Stacking proves to be justified in our case, as the whole ensemble attains AUC equal to 0.813, which is 0.026 higher than the best-performing base classifier, CatBoost (AUC = 0.787) and 0.043 better than LightGBM (AUC = 0.770). Our six synthetic features add another 0.011 of AUC to our model without any increase in the precision–recall curve.

Table 3. Ablation of model components and engineered features on the IBM test set.

Configuration	AUC	AUC-PR
Full ensemble (all features)	**0.813**	0.547
Ensemble without engineered features	0.802	**0.578**
CatBoost only	0.787	0.503
LightGBM only	0.770	0.452

F. Global and local explanations

Figure 5 shows the SHAP global summary plot. As time progresses, aspects like overtime, stock option values, age, monthly salary, and engineered mean-satisfaction values become the significant influencing factors for predicting employee attrition; however, one aspect which is especially important here is overtime, since overtime is seen to increase significantly the predicted propensity to leave, in alignment with the engineered overtime by low satisfaction interaction. The highest possible risk employee (who left and whose probability of leaving was 0.95) in the test dataset had his/her prediction attributed mainly due to high overtime, low monthly salary, and low satisfaction using local SHAP analysis. However, when looking at the generated DiCE counterfactuals, the predictions can be changed.

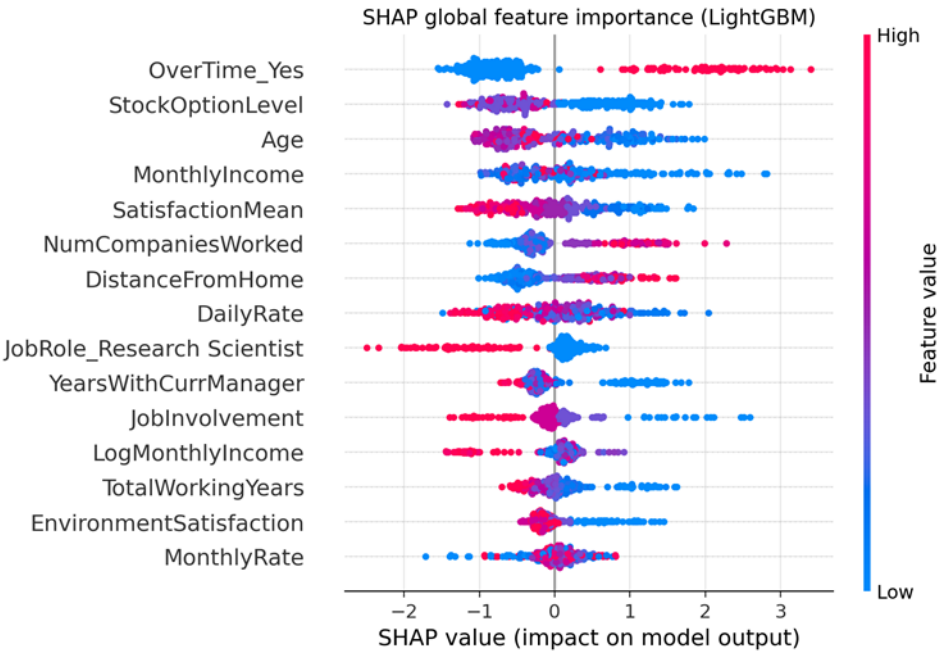

Fig. 5. SHAP global feature importance (beeswarm) for the LightGBM base learner on the IBM test set. Red = high feature value, blue = low.

G. Survival analysis

Fig. 6 shows Kaplan–Meier survival curves on the turnover dataset, overall and stratified by gender; the estimated median time-to-attrition is approximately 50.7 months. The Cox proportional-hazards model attains a concordance index of 0.66, with industry and

profession indicators among the strongest hazard contributors. The survival lens thus characterizes not only who leaves but when, information unavailable from a static classifier.

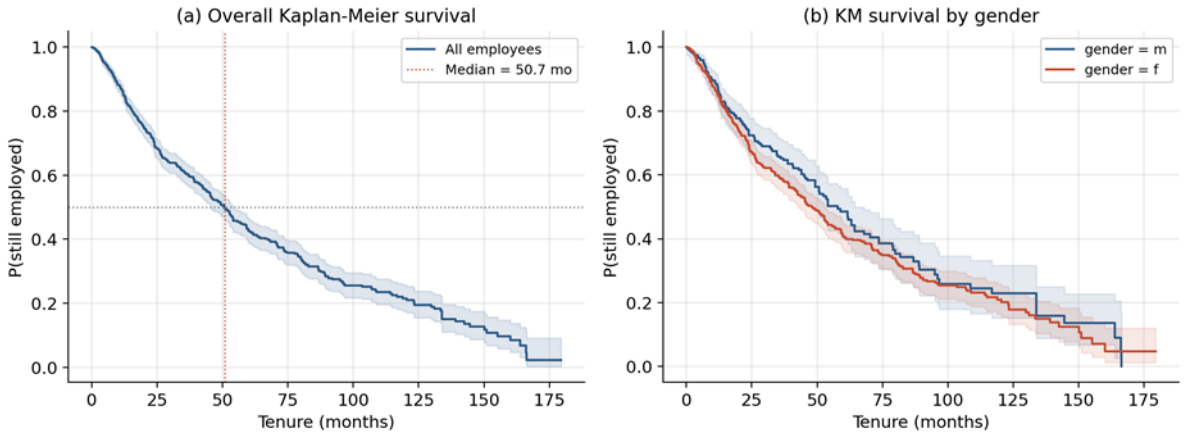

Fig. 6. Kaplan–Meier survival curves on the turnover dataset: (a) overall, with median survival marked; (b) stratified by gender, with 95% confidence bands.

H. Survival versus classification

The table below shows a controlled experiment using the same turnover holdout test. In terms of AUC relative to the event label, the classifier performs better (0.73 vs. 0.62). On the other hand, both models perform similarly with regards to concordance relative to actual durations (0.59 vs. 0.57). The takeaway is that for rank ordering where all that matters is whether the worker leaves or not, a classifier is enough. But if you care about either the time the worker leaves or hazard ratios, the Cox regression works well too.

Table 4. Survival versus classification on a common turnover hold-out split.

Model	AUC (event)	Concordance index
LightGBM classifier	**0.727**	**0.586**
Cox proportional hazards	0.623	0.574

I. Uplift estimation and the risk–uplift divergence

The three meta-learners produce moderately-to-strongly correlated uplift rankings (Spearman: S–T 0.72, T–X 0.79, S–X 0.56), indicating that the estimated responsiveness ordering is stable rather than an artifact of one estimator; we adopt the T-learner downstream. Fig. 7 plots each test employee by predicted risk and estimated uplift. The two dimensions are far from collinear: the top-50 risk set and the top-50 uplift set overlap with a Jaccard of just 0.27, so roughly three-quarters of the employees a risk-first policy would target differ from those a responsiveness-first policy would target. This divergence—rarely quantified in the attrition literature—motivates a targeting policy that reasons about responsiveness and cost rather than risk alone.

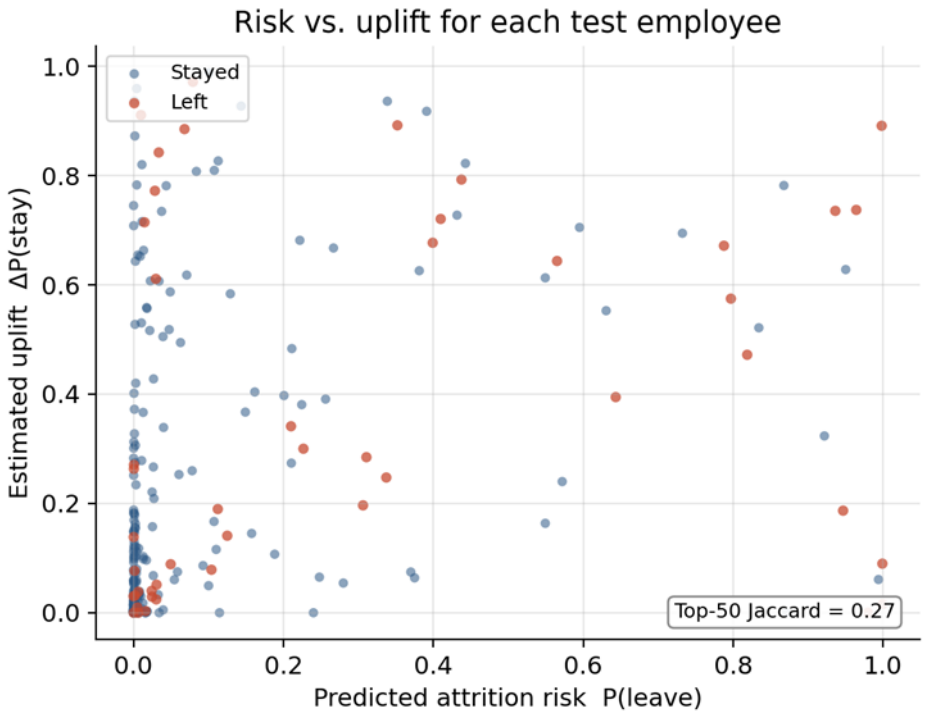

Fig. 7. Risk versus estimated uplift for each test employee. The low top-50 overlap shows risk-first and uplift-first targeting diverge substantially.

J. Fairness-Aware Cost-Sensitive Retention Targeting

Tables 5 and Figures 8 plot the proposed FACS-RT policy versus two baseline alternatives, namely risk-first and value-first, subject to an equal intervention budget of fifty. As compared to the other methods, the value-first method generates the highest expected value ($513k) but also produces the widest gender disparity (0.053). On the other hand, FACS-RT, with ε = 0.02, guarantees 83% of the expected value ($425k) while minimizing the disparity by 92% (0.004). Thus, FACS-RT attains the same level of fairness as the conservative risk-first policy but in much greater value terms. As seen in Figure 8(b), different values of ε trace the Pareto frontier of the value–fairness trade-off and present a clear elbow: the vast majority of disparity can be mitigated without a significant loss of value; however, beyond that threshold, further improvements are costly. Honest post-hoc evaluations using the ground-truth labels show that the risk-first policy has more leave-takers among the treatment group (21 out of 50), as opposed to the value-first (16 out of 50). This demonstrates once again the necessity of randomization for unbiased savings evaluation (see Section 7). The key insight of FACS-RT is the explicit governance of the value–fairness trade-off, where both boundary solutions correspond to the extreme choices of ε.

Table 5. Retention-targeting policies under a 50-intervention budget. Expected value is the ex-ante objective; realized true leavers and disparity are reported for validation.

Policy	Exp. value (US$k)	True leavers	Gender disparity
Risk-first	204	**21**	0.004
Value-first (= uplift-first)	**513**	16	0.053
FACS-RT (ε = 0.02)	425	10	**0.004**

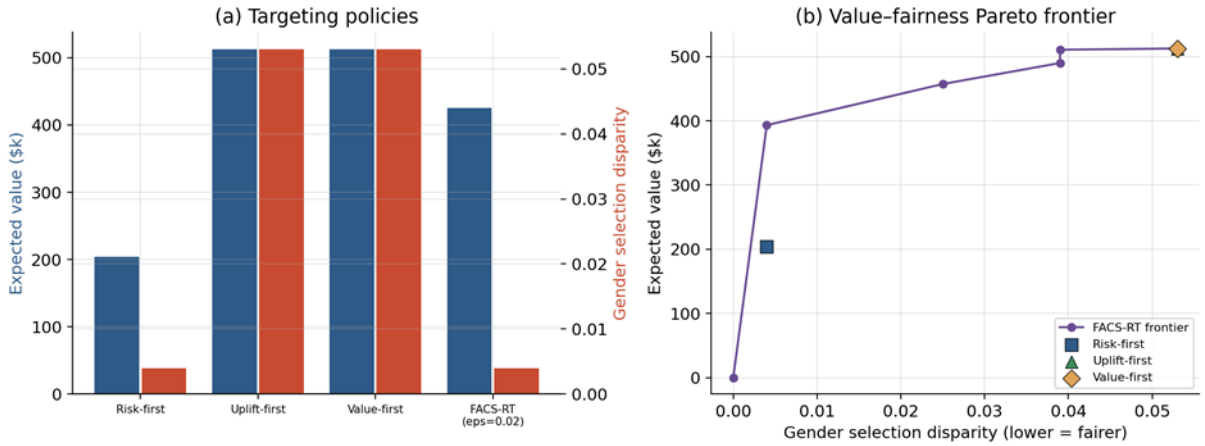

Fig. 8. FACS-RT. (a) Expected value (left axis) and gender selection disparity (right axis) by policy; (b) value–fairness Pareto frontier, with risk-first and value-first as boundary points.

K. Cross-domain generalization

Fig. 9 reports leave-one-department-out performance. The External AUC values span from 0.73 (Sales held-out) to 0.87 (Human Resources held-out), with the reference being 0.81 within domain. The discrepancy between AUC and AUC-PR is higher, reaching 0.48 for Sales and R&D since these domains comprise most examples belonging to the positive class, thus making precision more difficult to achieve due to distribution changes. The fluctuations indicate that the transferability of results from one dataset to another will be exaggerated, and a proper calibration is required before using the algorithm in an actual environment.

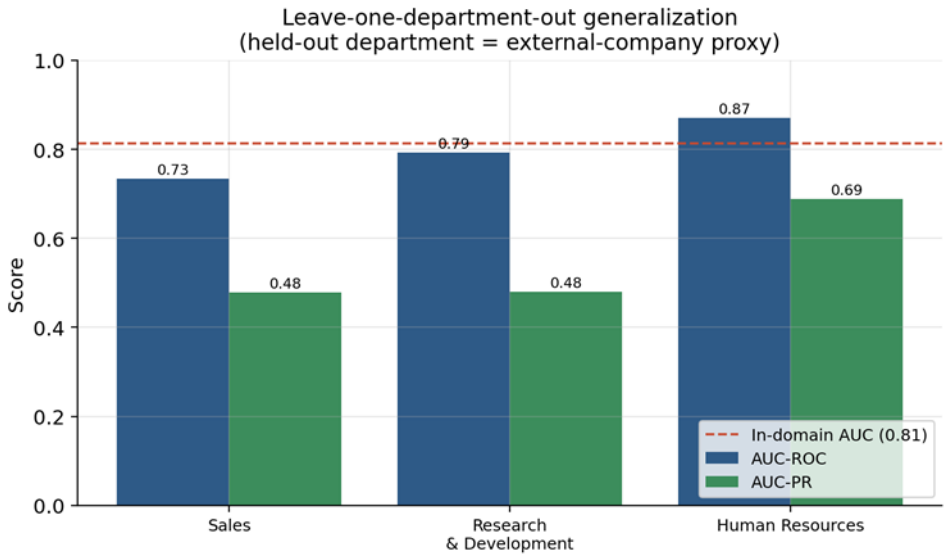

Fig. 9. Leave-one-department-out generalization. Bars show external AUC-ROC and AUC-PR for each held-out department; the dashed line is the in-domain AUC reference.

L. Second empirical base: turnover classification

To avoid making any conclusions based on a single dataset, we replicate the results on a different but similar dataset, where the target is a prediction of staff turnover. While the ranking of the models is similar to IBM results (enabling ensembles prevail), the actual regime of performance changes greatly since the target is now imbalanced (with 50.6% events); thus, the ensemble achieves the respective AUC value of 0.72 and F1 of 0.66, while logistic regression works much worse (with AUC 0.62) compared to the previous experiment. In conclusion, no generalizations can be made based on results obtained on only one HR dataset.

Table 6. Predictive performance on the turnover dataset (25% hold-out).

Model	AUC	AUC-PR	F1
Logistic Regression	0.618	0.625	0.590
Random Forest	**0.715**	**0.690**	**0.671**
XGBoost	0.687	0.661	0.623
Stacked Ensemble	**0.715**	0.674	0.657

M. Fairness audit

Table 7 shows group fairness metrics for the unconstrained ensemble (thus, before applying FACS-RT). There is no substantial disparity in the demographic parity metric across gender groups (0.038, lower than 0.05 threshold); nevertheless, equalized-odds discrepancy reaches 0.137, higher than the threshold. The fact that it is present implies different error rates between two groups, with women being more likely to get selected (0.30 against 0.26) and having better recall rate (0.75 against 0.61). This pattern persists in age groups too. The stated gap is the very essence of FACS-RT constraint.

Table 7. Group-fairness metrics for the unconstrained ensemble across gender.

Metric	Female	Male	Difference
Selection rate	0.302	0.264	0.038
Recall (true-positive rate)	0.750	0.613	**0.137**
Equalized-odds difference	—	—	**0.137**

V. CONCLUSION

We presented an integrated, decision-oriented framework for employee attrition that unifies cost-sensitive and calibrated prediction, dual-mode explainability, survival analysis, multi-learner uplift, fairness-constrained targeting, cross-domain validation, and statistical rigor across three public datasets, together with FACS-RT, a novel targeting policy that governs the value–fairness trade-off directly; the framework attains competitive, cross-validated discrimination (ensemble AUC = 0.83, 95% CI 0.79–0.89) while exposing what prediction-only studies miss—an ensemble statistically tied with a linear baseline, severe miscalibration that calibration repairs (ECE 0.22 → 0.04), a 27% risk–uplift overlap, cross-context AUC variability from 0.73 to 0.87, and a value–fairness frontier on which FACS-RT retains 83% of expected value while reducing gender selection disparity by 92%—and future work will validate the pipeline on a genuinely external, real-world HR dataset, replace the assumption-based uplift with a randomized retention experiment to obtain unbiased treatment effects and a label-faithful evaluation of FACS-RT savings, extend the survival track with machine-learning survival models such as random survival forests and gradient-boosted survival together with time-dependent covariates, and generalize FACS-RT to multi-attribute and intersectional fairness constraints and to in-processing optimization so that value and equity are optimized jointly rather than governed only at selection time.

VI. REFERENCES

1. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. DOI: 10.1023/A:1010933404324
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–

794). DOI: 10.1145/2939672.2939785 · arXiv: 1603.02754

3. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3146–3154). URL: proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

4. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* (Vol. 31, pp. 6638–6648). URL: proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html · arXiv: 1706.09516

5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. DOI: 10.1613/jair.953

6. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). URL: proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html · arXiv: 1705.07874

7. Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607–617). DOI: 10.1145/3351095.3372850

8. Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481. DOI: 10.1080/01621459.1958.10501452

9. Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–202. DOI: 10.1111/j.2517-6161.1972.tb00899.x

10. Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10), 4156–4165. DOI: 10.1073/pnas.1804597116

11. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732. DOI: 10.15779/Z38BG31 · SSRN: ssrn.com/abstract=2477899

12. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (Vol. 29, pp. 3315–3323). URL: proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html · arXiv: 1610.02413

13. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI* (Microsoft Technical Report MSR-TR-2020-32). Microsoft Research. URL: microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

14. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. URL: jmlr.org/papers/v12/pedregosa11a.html

15. Davidson-Pilon, C. (2019). lifelines: Survival analysis in Python. *Journal of Open Source Software*, 4(40), 1317. DOI: 10.21105/joss.01317

16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). DOI: 10.1145/2939672.2939778 · arXiv: 1602.04938

17. Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259. DOI: 10.1016/S0893-6080(05)80023-1

18. Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (pp. 973–978). URL: dl.acm.org/doi/10.5555/1642194.1642224 · PDF: cseweb.ucsd.edu/~elkan/rescale.pdf

19. Allen, D. G., Bryant, P. C., & Vardaman, J. M. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies. *Academy of Management Perspectives*, 24(2), 48–64. DOI:

10.5465/amp.24.2.48 · JSTOR: jstor.org/stable/25682398 *Note: corrected from your draft — no "Gusterson & Allen (2017)" paper with this title exists; the canonical reference is the 2010 Allen/Bryant/Vardaman paper above.*

20. Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360. DOI: 10.1073/pnas.1510489113 · arXiv: 1504.01132

21. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18), 2543–2546. DOI: 10.1001/jama.1982.03320430047030

22. IBM. (2017). *IBM HR Analytics Employee Attrition & Performance dataset*. Kaggle. URL: kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset