Anusandhanvallari
Vol 2024, No.1
November 2024
ISSN 2229-3388

Explainable and Fair Employee Attrition Prediction: A Cost-Sensitive Hr Analytics Approach

Prof. V. Lalitha

Professor Aradhana School of Business Management Hyderabad, Telangana, India

Abstract: Organisations increasingly use predictive HR analytics to identify employees at risk of leaving. Much of this work, however, reports only classification accuracy, neglecting whether complex models outperform simple ones, what drives their predictions, how predictions should be converted into decisions, and whether they treat employee groups fairly. This study addresses these four questions within a single, transparent analytical framework. Using the public IBM HR Analytics Employee Attrition dataset (N = 1,470; 30 predictors; 16.1% attrition), eight classifiers were compared through 5-fold cross-validation repeated three times, with differences tested by the corrected resampled t-test and the Friedman test. Logistic regression, random forest and gradient boosting were then tuned and evaluated on a stratified hold-out set of 441 employees. Regularised logistic regression achieved a cross-validated AUC of 0.835, statistically outperforming random forest (0.799), support vector machine (0.806), decision tree, k-nearest neighbours and naive Bayes, and matching gradient boosting (0.806 untuned; 0.836 tuned). On the hold-out set, the selected class-weighted logistic model achieved an AUC of 0.817 (95% CI 0.758–0.874) and captured 43.7% of leavers within the top 10% of risk scores. Overtime was the dominant predictor across all models, followed by years since promotion, job-hopping history, job satisfaction and business travel. A cost-optimal threshold of 0.62 reduced expected attrition cost by 12.8% relative to no intervention under base assumptions. The fairness audit revealed that employees under 30 and single employees were flagged at substantially higher rates and that true-positive rates differed by up to 22 percentage points across age groups. The study shows that interpretable models can match complex machine learning in HR prediction, and it offers an end-to-end template linking prediction, explanation, economics and fairness for responsible people analytics.

Keywords: HR analytics; People analytics; Employee attrition; Machine learning; Logistic regression; Gradient boosting; Explainable AI; Cost-sensitive learning; Algorithmic fairness; Human resource management

1. Introduction

Human resource management has long been criticised for relying on intuition and tradition rather than evidence. The emergence of **HR analytics**, also called people analytics or workforce analytics, promises to change this by applying statistical and computational methods to employee data in order to improve decisions about hiring, development, performance and retention (Davenport et al., 2010; Marler & Boudreau, 2017). Among its many applications, predicting employee attrition has become one of the most popular, because voluntary turnover is costly and because organisations already hold rich records on their employees. Yet the promise of HR analytics has not been fully realised. Reviews of the field note that much practice remains descriptive and that HR functions often lack the analytical capability to move beyond reporting (Angrave et al., 2016; Marler & Boudreau, 2017). At the same time, the growing use of machine learning raises new concerns. Complex algorithms may be difficult to explain to managers and employees; their predictions must be translated into decisions whose costs and benefits are rarely made explicit; and models trained on historical data may treat some groups of employees unfairly (Tambe et al., 2019). Studies of attrition prediction frequently report only the accuracy of a single model, a metric that can be misleading when, as is typical, only a minority of employees leave. The study uses the publicly available IBM HR Analytics dataset, which allows complete transparency and replication. Its contributions are threefold.

2. Literature Review

2.1 HR Analytics: Promise and Challenges

HR analytics has been defined as an HR practice enabled by information technology that uses descriptive, visual and statistical analyses of data related to HR processes, human capital, organisational performance and external benchmarks to establish business impact and enable data-driven decision-making (Marler & Boudreau, 2017). Davenport et al. (2010) described a progression from basic human capital reporting through analytical HR, value models and predictive talent models to real-time talent supply-chain optimisation. Levenson (2011) argued that analytics should be targeted at specific talent decisions with clear business value rather than at producing large volumes of metrics. Several authors have warned that HR analytics could become a management fad unless it is tied to business outcomes and grounded in sound methods (Rasmussen & Ulrich, 2015). Angrave et al. (2016) observed that many HR departments lack the skills to ask the right questions of data and that analytics vendors often emphasise dashboards over insight. Huselid (2018) called for workforce analytics to focus on strategic positions and outcomes. Tambe et al. (2019) identified four challenges specific to applying artificial intelligence in HR: the complexity of HR outcomes, the small size of many HR datasets, accountability for decisions that affect people's livelihoods, and the possible reactions of employees to algorithmic management. Each of these challenges is relevant to attrition prediction.

2.2 Theoretical Drivers of Attrition

Predictive models should be informed by theory about why people leave. Turnover research identifies job attitudes, particularly satisfaction and commitment, as proximal antecedents of quitting, alongside perceived alternatives, workload, compensation, career opportunities and supervisory relationships (Griffeth et al., 2000; Hom et al., 2017). Turnover is costly not only directly but also through its effect on organisational performance (Hancock et al., 2013). A predictive model that identifies the same drivers as theory gains credibility with decision-makers; one that relies on proxies without theoretical meaning should be treated with caution.

2.3 Machine Learning for Attrition Prediction

Machine learning methods differ from traditional statistical models in their emphasis on predictive accuracy rather than parameter interpretation (Breiman, 2001b). Common algorithms for binary classification include logistic regression, decision trees, ensemble methods such as random forests (Breiman, 2001a) and gradient boosting (Friedman, 2001), support vector machines, k-nearest neighbours and naive Bayes (Hastie et al., 2009). Ensemble methods often perform well on structured data because they capture non-linear relationships and interactions automatically.

However, the advantage of complex models is not guaranteed. On small, tabular datasets with modest signal, a well-regularised linear model may perform as well as or better than more flexible alternatives, and it offers far greater transparency (Rudin, 2019). Rigorous comparison therefore requires resampling-based evaluation and appropriate statistical tests. Because the folds of repeated cross-validation share training data, naive paired t-tests overstate the significance of differences; Nadeau and Bengio (2003) proposed a corrected resampled t-test that accounts for this dependence, and Demšar (2006) recommended non-parametric tests such as the Friedman test for comparing multiple classifiers.

2.4 Evaluating Imbalanced Classification

Because only a minority of employees leave, accuracy is a misleading metric: a model that predicts that nobody leaves would be correct for about 84% of employees in a typical dataset. The area under the receiver operating characteristic curve (AUC) measures a model's ability to rank leavers above stayers regardless of threshold (Fawcett, 2006). When the positive class is rare, the precision–recall curve and its summary, average precision, are more informative about performance in the region that matters (Saito & Rehmsmeier, 2015). The Brier score

evaluates the accuracy of predicted probabilities, and calibration plots show whether predicted risks match observed rates (Niculescu-Mizil & Caruana, 2005).

2.5 From Prediction to Decision

A probability is not a decision. Converting predicted risk into action requires a threshold, and the optimal threshold depends on the relative costs of different errors (Elkan, 2001). In attrition management, failing to identify an employee who later leaves (a false negative) incurs the full replacement cost, while flagging an employee who would have stayed (a false positive) incurs the cost of an unnecessary retention intervention. The default threshold of 0.5 is rarely optimal under such asymmetric costs. Cumulative gains or lift analysis shows how efficiently a model concentrates leavers among those it ranks as highest risk, which is directly relevant when retention resources are limited.

2.6 Explainability and Fairness

Managers and employees are unlikely to trust, or act on, predictions they cannot understand. Model-agnostic explanation methods help. **Permutation importance** measures how much a model's performance deteriorates when the values of a feature are randomly shuffled (Breiman, 2001a; Fisher et al., 2019). **Partial dependence** plots show the average predicted outcome as a function of one feature, marginalising over the others (Friedman, 2001). Shapley-value-based methods provide local explanations for individual predictions (Lundberg & Lee, 2017). Dr. Naveen Prasadula (2024) reviews these approaches.

Algorithmic fairness has become a central concern wherever predictions affect people. Several criteria have been proposed: **demographic parity** requires similar selection rates across groups; **equal opportunity** requires similar true-positive rates (Hardt et al., 2016); and **predictive parity** requires similar precision. These criteria generally cannot all be satisfied simultaneously when base rates differ between groups (Chouldechova, 2017; Kleinberg et al., 2017), so organisations must choose which notion of fairness matters most in a given context (Barocas et al., 2023). Toolkits such as IBM's AI Fairness 360 support such audits (Bellamy et al., 2019). In employment, the "four-fifths rule", under which a selection rate for any group below 80% of the highest group's rate is treated as evidence of adverse impact, is widely used as a practical benchmark. In India, the Digital Personal Data Protection Act, 2023 adds obligations regarding the processing of employees' personal data.

2.7 Research Gap

Existing studies of attrition prediction tend to focus on one element, usually accuracy, of what is in practice a chain of analytical steps. Few combine rigorous model comparison with explanation, economic decision analysis and fairness assessment. This study fills that gap.

3. Methodology

3.1 Data

The study uses the IBM HR Analytics Employee Attrition and Performance dataset (Subhash, 2017), created by IBM data scientists and distributed under the Open Database License, including through the IBM employee-attrition-aif360 repository (IBM, n.d.). It contains 1,470 employee records with 35 variables. Four variables were removed because they are constant (employee count, over-18 flag, standard hours) or an identifier (employee number), leaving 30 predictors: 23 numeric and 7 categorical. The outcome is attrition (1 = left; 0 = stayed), with 237 leavers (16.1%). Because the data are fictional, no personal information is involved.

The predictors cover demographics (age, gender, marital status, education, education field, distance from home), job characteristics (department, job role, job level, business travel, overtime), compensation (monthly income, daily, hourly and monthly rates, percentage salary hike, stock option level), career history (total working years, number of previous companies, years at company, years in current role, years since last promotion, years with

current manager, training sessions) and attitudes (job satisfaction, environment satisfaction, relationship satisfaction, job involvement, work-life balance) together with the last performance rating.

Table 1 profiles the sample and compares employees who stayed with those who left on selected variables, to show the raw patterns that the models must learn.

Table 1: Sample profile by attrition status (N = 1,470)

Variable	All employees	Stayed (n = 1,233)	Left (n = 237)	Test of difference
Age (years), mean	36.9	37.6	33.6	$t = 5.83$, $p < .001$
Monthly income, mean	6,503	6,833	4,787	$t = 7.48$, $p < .001$
Total working years, mean	11.3	11.9	8.2	$t = 7.02$, $p < .001$
Years with current manager, mean	4.1	4.4	2.9	$t = 6.63$, $p < .001$
Distance from home (km), mean	9.2	8.9	10.6	$t = -2.89$, $p = .004$
Job satisfaction (1–4), mean	2.73	2.78	2.47	$t = 3.93$, $p < .001$
Job involvement (1–4), mean	2.73	2.77	2.52	$t = 4.66$, $p < .001$
Works overtime (%)	28.3	23.4	53.6	$\chi^2 = 87.56$, $p < .001$
Single (%)	32.0	28.4	50.6	$\chi^2 = 46.16$, $p < .001$
Travels frequently (%)	18.8	16.9	29.1	$\chi^2 = 24.18$, $p < .001$
Female (%)	40.0	40.6	36.7	$\chi^2 = 1.12$, $p = .291$

Note: t = Welch t-test; χ^2 = Pearson chi-square test on the full categorical variable (overtime: 2 categories; marital status and business travel: 3 categories).

Leavers were younger, earned less, had shorter careers and shorter relationships with their managers, lived further from work and reported lower job satisfaction and involvement. More than half of leavers worked overtime, compared with fewer than a quarter of stayers. Gender was not associated with attrition.

3.2 Pre-processing and Data Partitioning

Numeric predictors were standardised to zero mean and unit variance, and categorical predictors were one-hot encoded (binary variables with a single indicator). All pre-processing was embedded in a pipeline fitted only on training data within each fold, preventing information from validation or test data leaking into model training. The data were split into a training set of 1,029 employees (70%) and a hold-out test set of 441 employees (30%),

stratified by attrition so that both sets contain 16.1% leavers. The test set was used only once, for final evaluation.

3.3 Models

Eight classifiers were compared:

- **Logistic regression** with L2 regularisation, in standard and class-weighted forms (the latter giving more weight to the minority class).
- **Decision tree** with a maximum depth of five and a minimum of ten observations per leaf.
- **Random forest** with 500 trees and balanced class weights (Breiman, 2001a).
- **Gradient boosting** with default settings (Friedman, 2001).
- **Support vector machine** with a radial basis function kernel and balanced class weights.
- **k-nearest neighbours** with k = 15.
- **Gaussian naive Bayes.**

3.4 Evaluation Design

Models were compared on the training set using 5-fold stratified cross-validation repeated three times, giving 15 performance estimates per model. The primary metric was AUC; secondary metrics were average precision (AP), F1 score, recall and balanced accuracy at the default 0.5 threshold, and the Brier score. Differences between the best model and each alternative were tested with the corrected resampled t-test (Nadeau & Bengio, 2003), and differences among all eight with the Friedman test (Demšar, 2006).

The three most relevant model families (class-weighted logistic regression, random forest and gradient boosting) were then tuned by grid search with 5-fold cross-validation on the training set, maximising AUC. The final model was selected by a parsimony rule fixed in advance: among models whose tuned cross-validated AUC was within 0.01 of the best, the most interpretable was chosen. All three tuned models were evaluated on the test set, with 95% confidence intervals for AUC obtained from 2,000 bootstrap resamples.

3.5 Explainability

Permutation importance was computed on the test set as the mean decrease in AUC over 30 random permutations of each predictor, for each of the three tuned models. Agreement among models was summarised by Spearman rank correlations of importance scores. Standardised logistic coefficients were reported as odds ratios. Partial dependence functions were estimated from the tuned gradient boosting model for six key numeric predictors.

3.6 Cost-Sensitive Threshold Selection

An expected-cost model was specified. The cost of losing an employee is normalised to 1 unit (approximately the cost of replacing the employee). A retention intervention (for example a stay interview, workload adjustment, development plan or targeted reward) costs C_I = 0.10 units and succeeds in retaining a would-be leaver with probability s = 0.40. Expected cost per employee for a threshold t is therefore:

$$ Cost(t) = [FN + TP × (C_I + (1 – s)) + FP × C_I] / N $$

where TP, FP and FN are counts at threshold t. The baseline of no intervention costs the attrition rate (0.161 units per employee). The optimal threshold minimises Cost(t). Because these parameters vary between organisations, a sensitivity analysis varied C_I from 0.05 to 0.30 and s from 0.25 to 0.60.

3.7 Fairness Audit

At the cost-optimal threshold, selection rate (proportion flagged), true-positive rate (recall), false-positive rate, precision and group-specific AUC were compared across gender, age group (under 30 vs 30 and above) and marital status. Disparities were assessed against the four-fifths rule for selection rates and against the equal-opportunity criterion for true-positive rates.

3.8 Software

Analyses were performed in Python using pandas, NumPy, SciPy and scikit-learn (Pedregosa et al., 2011), with a fixed random seed for reproducibility.

4. Results

4.1 Cross-Validated Model Comparison

Table 2 reports cross-validated performance. Standard logistic regression achieved the highest mean AUC (0.835) and average precision (0.628), followed by class-weighted logistic regression (AUC 0.824). Gradient boosting and the support vector machine (both 0.806) and random forest (0.799) performed somewhat worse, while k-nearest neighbours (0.724) and the decision tree (0.703) performed poorest. The Friedman test confirmed significant differences among the eight models ($\chi^2 = 84.07$, df = 7, $p < .001$).

Table 2: Cross-validated performance of eight classifiers (5-fold × 3 repeats; mean and SD)

Model	AUC	Average precision	F1 (0.5)	Recall (0.5)	Balanced accuracy	Brier score
Logistic regression	0.835 (0.034)	0.628 (0.054)	0.522	0.417	0.692	0.094
Logistic regression (weighted)	0.824 (0.036)	0.602 (0.056)	0.510	0.751	0.761	0.159
Gradient boosting	0.806 (0.034)	0.560 (0.066)	0.445	0.333	0.652	0.102
Support vector machine	0.806 (0.031)	0.568 (0.058)	0.434	0.317	0.646	0.102
Random forest	0.799 (0.041)	0.558 (0.070)	0.362	0.237	0.613	0.106
Naive Bayes	0.747 (0.040)	0.407 (0.089)	0.390	0.751	0.672	0.324
k-nearest neighbours	0.724 (0.036)	0.391 (0.065)	0.063	0.034	0.517	0.121
Decision tree	0.703 (0.030)	0.393 (0.043)	0.359	0.290	0.616	0.124

Note: Training set N = 1,029. Threshold-dependent metrics use a 0.5 cut-off. Lower Brier scores indicate more accurate probabilities.

Table 3 reports the corrected resampled t-tests comparing the best model, standard logistic regression, with each alternative. Logistic regression significantly outperformed the random forest (ΔAUC = 0.035, p = .038), the support vector machine (ΔAUC = 0.028, p = .036), naive Bayes, k-nearest neighbours and the decision tree. Its advantage over untuned gradient boosting (ΔAUC = 0.028) was marginal (p = .053).

Table 3: Corrected resampled t-tests: standard logistic regression versus alternatives (AUC)

Comparison model	Mean AUC difference	Corrected t (df = 14)	p
Logistic regression (weighted)	0.010	3.37	.005
Gradient boosting	0.028	2.11	.053
Support vector machine	0.028	2.31	.036
Random forest	0.035	2.29	.038
Naive Bayes	0.087	3.65	.003
k-nearest neighbours	0.110	4.06	.001
Decision tree	0.131	8.56	< .001

The threshold-dependent metrics highlight an important practical point. At the default 0.5 threshold, unweighted models identified only a minority of leavers (recall between 3% and 42%), whereas class-weighted logistic regression identified 75% of them, at the cost of more false alarms and less accurate probabilities (Brier score 0.159). The choice of threshold and weighting is therefore as important as the choice of algorithm.

4.2 Tuned Models and Hold-Out Performance

Grid search selected an inverse regularisation strength of C = 0.1 for class-weighted logistic regression (cross-validated AUC 0.831); unrestricted depth, square-root feature sampling and a minimum leaf size of one for the random forest (0.811); and 400 shallow trees (depth 2) with a learning rate of 0.1 and 80% subsampling for gradient boosting (0.836). Because the tuned logistic model was within 0.005 of the best cross-validated AUC, it was selected as the final model under the parsimony rule.

Table 4 shows hold-out performance. The tuned logistic regression achieved the highest test AUC (0.817; 95% CI 0.758–0.874) and average precision (0.601), ahead of gradient boosting (0.803) and random forest (0.780). The bootstrap difference in AUC between logistic regression and random forest was 0.037 (95% CI –0.009 to 0.081), and between logistic regression and gradient boosting 0.013 (95% CI –0.036 to 0.062); neither interval excludes zero, so the three models should be regarded as statistically comparable on this test set, with no evidence that the more complex models are superior.

Table 4: Hold-out test performance of tuned models (N = 441; threshold 0.5)

Metric	Logistic regression (weighted, tuned)	Random forest (tuned)	Gradient boosting (tuned)
AUC (95% CI)	0.817 (0.758–0.874)	0.780 (0.719–0.839)	0.803 (0.744–0.860)
Average precision	0.601	0.421	0.512
Brier score	0.152	0.116	0.111
Accuracy	0.773	0.841	0.862
Precision	0.386	0.533	0.656
Recall (sensitivity)	0.690	0.113	0.296
Specificity	0.789	0.981	0.970
F1 score	0.495	0.186	0.408
Balanced accuracy	0.740	0.547	0.633

The comparison again shows why accuracy alone misleads. The random forest had the second-highest accuracy (84.1%) but identified only 8 of 71 leavers (recall 11.3%). The weighted logistic model had lower accuracy (77.3%) but identified 49 of 71 leavers (69.0%).

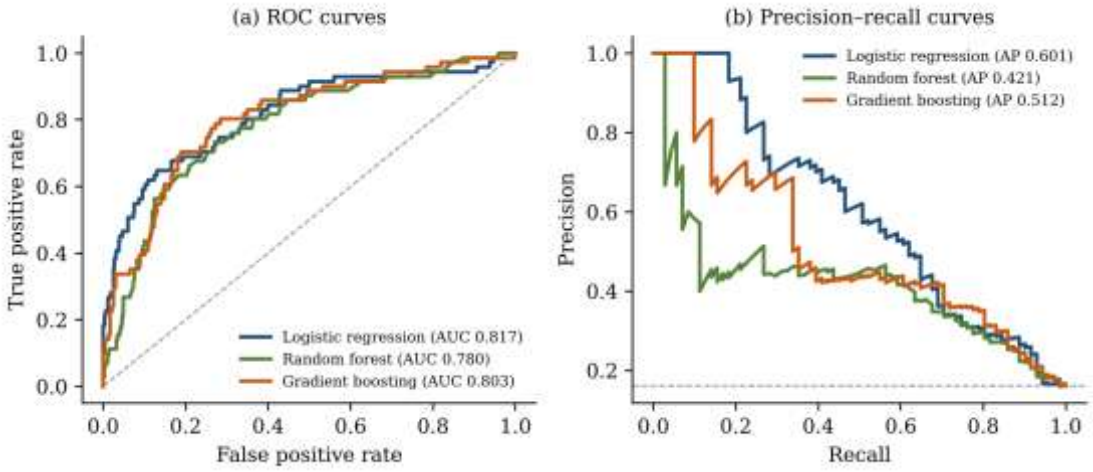


Figure 1: (a) ROC curves and (b) precision–recall curves for the three tuned models on the hold-out test set. The dashed line in (b) shows the attrition base rate of 16.1%.

Figure 1 displays the ROC and precision–recall curves. The logistic model's advantage is most visible in the precision–recall curve at low recall levels, that is, among the employees ranked as highest risk, where retention efforts would be concentrated.

4.3 Calibration

Class weighting improves recall at a fixed threshold but distorts predicted probabilities. Figure 2 shows that the weighted logistic model systematically over-predicts attrition: employees assigned a probability of about 0.5 leave at a much lower rate. The Brier score (0.152) is accordingly higher than for the unweighted ensembles (0.111–0.116). For ranking and threshold-based decisions this does not matter, but if the probabilities themselves are to be communicated, for example as "a 40% chance of leaving", recalibration or an unweighted model should be used.

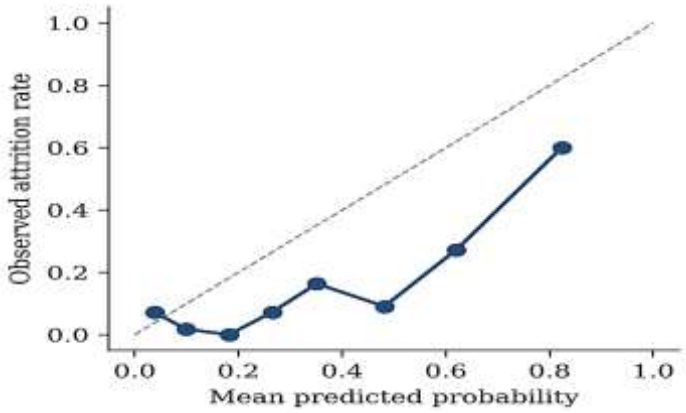


Figure 2: Calibration plot for the final model on the test set (eight quantile bins). Points below the diagonal indicate over-prediction.

4.4 What Drives Predicted Attrition?

Figure 3 presents permutation importance for the final model. Overtime was by far the most influential predictor: permuting it reduced test AUC by 0.077, nearly three times as much as any other variable. It was followed by years since last promotion (0.028), number of companies previously worked for (0.024), job satisfaction (0.024), business travel (0.018), marital status (0.015), monthly income (0.014), environment satisfaction (0.013) and job involvement (0.010).

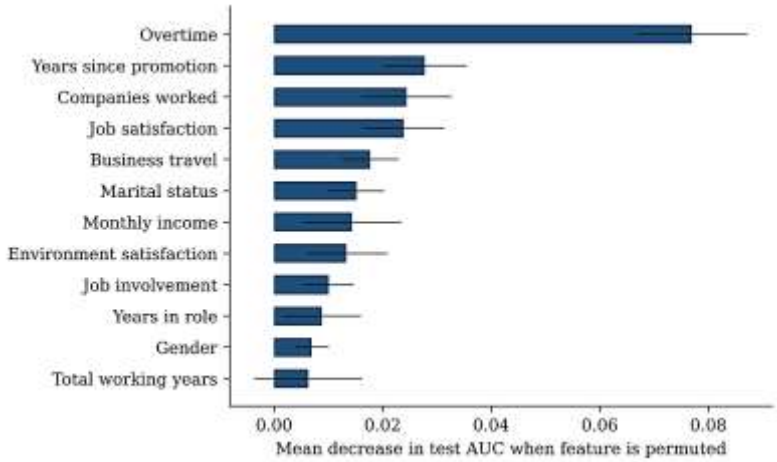


Figure 3: Permutation importance of the twelve most important predictors in the final logistic regression model (mean decrease in test AUC over 30 permutations; error bars show one standard deviation).

The drivers were broadly consistent across model families. Overtime ranked first for all three tuned models, with AUC decreases of 0.077 (logistic regression), 0.070 (random forest) and 0.094 (gradient boosting). Job satisfaction, job involvement, stock option level, business travel and income appeared among the top ten in at least two models. Spearman rank correlations of importance across all 30 predictors were 0.50 between logistic regression and random forest, 0.60 between logistic regression and gradient boosting, and 0.67 between the two ensembles, indicating moderate agreement: the models agree on the main drivers but differ on the ordering of weaker ones.

Table 5 reports the largest standardised coefficients of the final logistic model as odds ratios per standard deviation (or relative to the reference category for dummy variables).

Table 5: Largest standardised logistic regression effects in the final model

Predictor	Coefficient	Odds ratio	Interpretation
Overtime (yes)	1.211	3.36	Overtime more than triples the odds
Job role: laboratory technician	0.599	1.82	Higher risk than reference role
Frequent business travel	0.545	1.73	Frequent travel raises risk
Years since last promotion (+1 SD)	0.478	1.61	Career stagnation raises risk
Job role: sales representative	0.430	1.54	Higher risk than reference role
Marital status: single	0.417	1.52	Single employees at higher risk
Number of companies worked (+1 SD)	0.398	1.49	Job-hopping history raises risk
Distance from home (+1 SD)	0.258	1.29	Longer commute raises risk
Monthly income (+1 SD)	–0.316	0.73	Higher pay protects
Environment satisfaction (+1 SD)	–0.321	0.73	Better environment protects
Total working years (+1 SD)	–0.337	0.71	Experience protects
Job satisfaction (+1 SD)	–0.347	0.71	Satisfaction protects
No business travel	–0.519	0.60	Not travelling protects

Note: Coefficients are from the tuned, class-weighted, L2-regularised model on standardised predictors; regularisation shrinks coefficients towards zero, so magnitudes are conservative.

4.5 Shape of the Relationships

Figure 4 shows partial dependence functions from the gradient boosting model. Predicted attrition fell from about 0.20 at the lowest income levels to around 0.10 at the highest, with most of the decline at the lower end and a secondary bump in the middle of the range that probably reflects sparse data there. Risk was highest for the least experienced employees (about 0.25 with one year of total experience) and declined to around 0.14 for those with long careers. Age showed a similar decline from early career, from about 0.21 in the early twenties to

about 0.12 in the late thirties. Risk rose steadily with commuting distance, from about 0.12 for those living closest to around 0.20 for those living furthest away, and more than doubled with the number of previous employers, from about 0.11 to 0.24. By contrast, tenure at the current company showed an almost flat relationship (0.15–0.18) once other variables were held constant, suggesting that the well-known early-tenure effect is largely explained by age, experience and income.

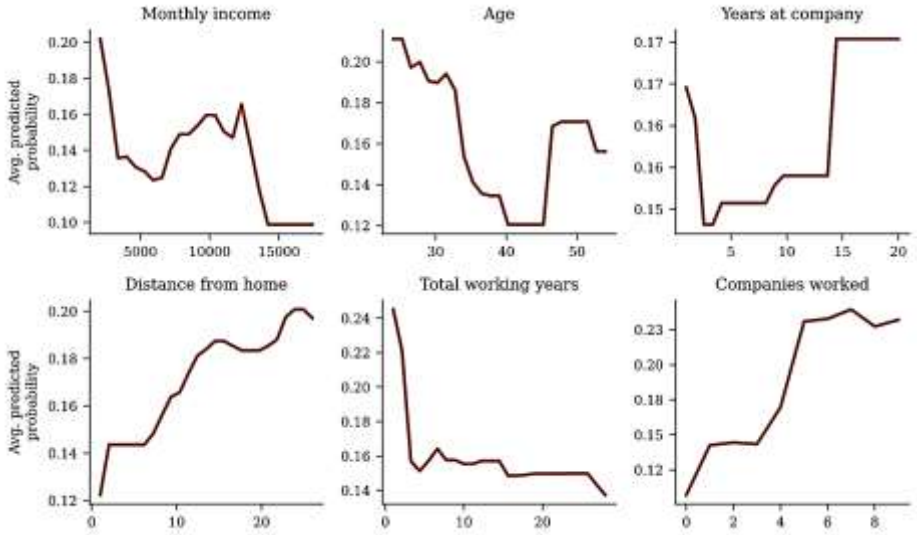


Figure 4: Partial dependence of predicted attrition probability on six numeric predictors (tuned gradient boosting model).

4.6 From Prediction to Decision

Figure 5(a) plots expected cost per employee against the classification threshold under base assumptions (intervention cost 0.10, success rate 40%). Without any intervention, expected cost equals the attrition rate, 0.161 units per employee. The cost curve reaches its minimum of 0.140 at a threshold of 0.62, a reduction of 12.8%. Using the default threshold of 0.5 gives 0.145.

At the optimal threshold, the model flags 81 of 441 employees (18.4%), of whom 43 are actual leavers (precision 53.1%), capturing 60.6% of all leavers while incorrectly flagging only 10.3% of stayers (specificity 89.7%). Accuracy rises to 85.0% and F1 to 0.566.

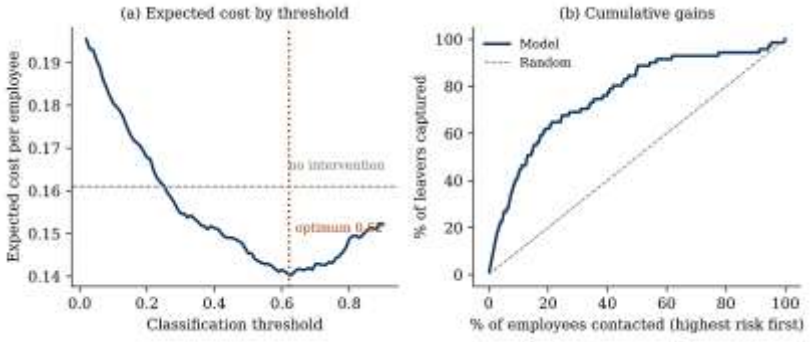


Figure 5: (a) Expected cost per employee by classification threshold, with the no-intervention baseline; (b) cumulative gains chart for the final model on the test set.

Figure 5(b) shows the cumulative gains chart. Contacting the 10% of employees with the highest predicted risk would reach 43.7% of all leavers, more than four times what random selection would achieve. The top 20% captures 62.0% of leavers, the top 30% captures 69.0%, and the top half captures 87.3%. For HR teams with limited capacity for stay interviews or retention offers, this concentration is the model's most tangible benefit.

Table 6: Sensitivity of the optimal threshold and cost saving to intervention assumptions

Intervention cost (C_I)	Success rate 25%	Success rate 40%	Success rate 60%
0.05	t = 0.62; saving 9.4%	t = 0.30; saving 19.7%	t = 0.30; saving 37.5%
0.10	t = 0.73; saving 4.6%	t = 0.62; saving 12.8%	t = 0.59; saving 25.1%
0.20	t = 0.90; saving 0.9%	t = 0.73; saving 4.8%	t = 0.69; saving 14.1%
0.30	t = 0.90; saving – 0.9%	t = 0.90; saving 1.8%	t = 0.73; saving 7.2%

Note: t = cost-minimising threshold; saving = reduction in expected cost relative to no intervention. Costs are expressed as a fraction of the cost of losing one employee.

Table 6 shows that the value of the model depends heavily on the economics of intervention. When interventions are cheap and effective (cost 5% of replacement cost, 60% success), the model reduces expected attrition costs by 37.5% and it pays to cast a wide net with a low threshold of 0.30. When interventions are expensive and rarely succeed (cost 30%, success 25%), the model offers no benefit: no intervention is better. The lesson for practice is that the quality of retention interventions matters at least as much as the quality of the prediction.

4.7 Fairness Audit

Table 7 reports the fairness audit at the cost-optimal threshold.

Table 7: Fairness audit of the final model at the optimal threshold (test set)

Group	n	Base rate	Selection rate	True-positive rate	False-positive rate	Precision	AUC
Female	169	15.40%	18.90%	65.40%	10.50%	53.10%	0.772
Male	272	16.50%	18.00%	57.80%	10.10%	53.10%	0.843
Under 30	95	25.30%	29.50%	75.00%	14.10%	64.30%	0.864
30 and above	346	13.60%	15.30%	53.20%	9.40%	47.20%	0.787
Single	146	24.70%	34.90%	69.40%	23.60%	49.00%	0.81
Married	196	13.80%	10.70%	40.70%	5.90%	52.40%	0.724
Divorced	99	8.10%	9.10%	87.50%	2.20%	77.80%	0.966

Gender. The model treated women and men similarly. Selection rates were close (18.9% and 18.0%; ratio 0.95, well within the four-fifths rule), precision was identical (53.1%) and false-positive rates were almost equal. The true-positive rate was 7.6 percentage points higher for women, a modest difference in women's favour from a retention perspective.

Age. Employees under 30 were flagged nearly twice as often as older employees (29.5% vs 15.3%; ratio 0.52, failing the four-fifths rule). This largely reflects a genuine difference in base rates (25.3% vs 13.6%), and precision was actually higher for younger employees. However, the true-positive rate differed by 21.8 percentage points (75.0% vs 53.2%), meaning that older employees who were about to leave were considerably less likely to be identified and offered support. This violates equal opportunity.

Marital status. Single employees were flagged at 34.9%, compared with 10.7% of married employees, and had a false-positive rate four times higher (23.6% vs 5.9%). Conversely, married employees who later left were identified only 40.7% of the time. Marital status is a legally and ethically sensitive characteristic, and its use as a predictor warrants scrutiny.

These patterns illustrate the impossibility results of Chouldechova (2017) and Kleinberg et al. (2017): when base rates differ, a single threshold cannot equalise selection rates, error rates and precision simultaneously. Which criterion to prioritise is a value judgement. For a supportive retention programme, equal opportunity (ensuring that leavers in every group have a similar chance of receiving help) is arguably the most relevant. For any use that could disadvantage flagged employees, such as excluding them from promotion or training, selection-rate parity and the exclusion of sensitive attributes become much more important.

4.8 Workforce Segmentation

As a complementary unsupervised analysis, k-means clustering was applied to twelve standardised work-experience variables (five attitudes, overtime, income, tenure, years since promotion, training, distance and age). The silhouette coefficient was highest for two clusters (0.18), indicating weak but discernible structure. The first segment (n = 333) consisted of established, senior employees (mean age 45.3 years, mean tenure 13.9 years, mean monthly income 12,824) and had an attrition rate of 10.0%. The second, much larger segment (n = 1,137) consisted of early- and mid-career employees (mean age 34.5 years, tenure 5.0 years, income 4,652) with an attrition rate of 18.0% ($\chi^2 = 12.89$, df = 1, $p < .001$). Attitudes and overtime were nearly identical in the two segments, confirming that attitudinal and workload risks cut across career stages, whereas career stage itself shifts baseline risk.

5. Discussion

5.1 Simplicity Can Win

The most important finding for HR analytics practice is that a regularised logistic regression matched or outperformed random forests, gradient boosting and support vector machines. In cross-validation it significantly outperformed most alternatives; on the hold-out set it achieved the highest AUC and average precision, although differences from gradient boosting were not statistically significant. This supports Rudin's (2019) argument that for many high-stakes decisions on structured data, interpretable models can perform as well as black boxes. HR datasets are typically small, noisy and dominated by a handful of strong predictors, conditions under which complex models gain little and risk overfitting. Practitioners should therefore begin with transparent models and adopt complex ones only when they demonstrate a clear, statistically reliable advantage.

5.2 Workload and Career Progression as Primary Levers

Across all models, overtime was the dominant driver of predicted attrition, followed by career stagnation, a history of frequent job changes, low job satisfaction and frequent business travel. These findings are consistent

with turnover theory, which emphasises job demands, satisfaction and perceived alternatives (Griffeth et al., 2000; Hom et al., 2017). Several of these drivers are directly controllable by the organisation: workload, travel requirements, promotion pacing and the work environment. Others, such as job-hopping history, are informative for risk prediction but are not levers for intervention, and should be used cautiously in hiring decisions to avoid penalising candidates unfairly.

5.3 The Economics of Prediction

The cost analysis shows that a model's value cannot be judged from its AUC alone. Under plausible assumptions the model reduced expected attrition costs by about 13%, but this ranged from nothing to more than a third depending on the cost and effectiveness of retention actions. Organisations should therefore invest in designing and evaluating effective, low-cost retention interventions before investing in more sophisticated prediction. The cumulative gains analysis offers a practical operating rule: with capacity to engage one in ten employees, a model like this reaches over 40% of those who would otherwise leave.

5.4 Fairness Is a Design Choice

The audit revealed that the model's errors are unevenly distributed across age and marital-status groups, even though gender parity was broadly satisfied. These disparities arise not from malicious design but from differences in base rates and from the use of predictors correlated with group membership. Organisations deploying such models should audit them routinely, decide explicitly which fairness criterion matters for each use, consider removing sensitive attributes such as marital status, use group-specific thresholds or post-processing where legally permissible, and ensure that predictions are used only to offer support, never to penalise.

5.5 Implications for HR Practice

- **Start simple and test rigorously.** Use cross-validation, appropriate metrics such as AUC and average precision, and corrected statistical tests before adopting complex models.
- **Report the right metrics.** Replace accuracy with recall, precision, average precision and gains at realistic contact rates.
- **Set thresholds on economic grounds.** Estimate the cost of losing an employee and the cost and success rate of interventions, and choose the threshold that minimises expected cost.
- **Explain predictions.** Share the main drivers with managers in plain language, and link each flagged case to actionable factors such as workload or time since promotion.
- **Act on the drivers, not just the scores.** Manage overtime, rationalise travel, maintain career progression and improve job satisfaction across the workforce, not only for flagged individuals.
- **Audit for fairness and protect privacy.** Monitor selection and error rates by group, exclude inappropriate attributes, comply with data protection law such as India's Digital Personal Data Protection Act, 2023, and be transparent with employees about how their data are used.
- **Use predictions to support, not to sanction.** Attrition risk scores should trigger supportive conversations, never adverse employment decisions.

6. Limitations and Future Research

Several limitations qualify these findings. First, the dataset is fictional, created to resemble real HR data; results should be validated on actual organisational records, which are typically larger, messier and more varied. Second, the data are cross-sectional, so the analysis predicts status rather than future departure; a deployment

model should use predictors measured before the outcome period and be validated on later cohorts. Third, the sample of 1,470 employees, with only 71 leavers in the test set, limits statistical power, which is reflected in wide bootstrap confidence intervals. Fourth, the cost model uses illustrative parameters; organisations should estimate their own. Fifth, the fairness audit covered only the attributes available and used a single threshold; intersectional analysis and mitigation methods deserve further study. Finally, explanation methods such as permutation importance can be affected by correlated predictors, and causal conclusions should not be drawn from predictive importance.

Future research could apply the framework to real data from Indian organisations in sectors such as IT services, banking and healthcare; use survival analysis to predict the timing of departure; incorporate text data from engagement surveys and exit interviews; evaluate retention interventions experimentally to estimate their success rates; and compare fairness mitigation techniques in HR settings.

7. Conclusion

This study developed and demonstrated an end-to-end framework for predictive HR analytics, combining rigorous model comparison, explanation, cost-sensitive decision-making and fairness auditing. Applied to the IBM HR Analytics dataset, it showed that an interpretable logistic regression performs as well as complex machine learning models; that overtime, career stagnation, job-hopping history, job satisfaction and business travel are the principal drivers of attrition risk; that a cost-optimal threshold can reduce expected attrition costs, but only when retention interventions are sufficiently cheap and effective; and that a model which appears fair by gender may still distribute its benefits unevenly by age and marital status.

The broader message is that the value of HR analytics lies not in the sophistication of algorithms but in the quality of the decisions they inform. Organisations that combine transparent models with sound economics, attention to fairness and genuine investment in employee well-being are best placed to turn data into retention.

Data Availability Statement

The IBM HR Analytics Employee Attrition dataset is publicly available from Kaggle and from the IBM employee-attrition-aif360 repository on GitHub under the Open Database License. The analysis code (Python, scikit-learn) is available from the corresponding author on request. A fixed random seed (42) was used for data partitioning, cross-validation and model training so that all results can be reproduced.

Conflict of Interest

The authors declare no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

[1] Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. Human Resource Management Journal, 26(1), 1–11.

[2] Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.

[3] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M.,

Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1–4:15.

[4] Breiman, L. (2001a). Random forests. Machine Learning, 45(1), 5–32.

[5] Breiman, L. (2001b). Statistical modeling: The two cultures. Statistical Science, 16(3), 199–231.

[6] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.

[7] Davenport, T. H., Harris, J., & Shapiro, J. (2010). Competing on talent analytics. Harvard Business Review, 88(10), 52–58.

[8] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30.

[9] Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (pp. 973–978).

[10] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

[11] Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20(177), 1–81.

[12] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

[13] Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. Journal of Management, 26(3), 463–488.

[14] Hancock, J. I., Allen, D. G., Bosco, F. A., McDaniel, K. R., & Pierce, C. A. (2013). Meta-analytic review of employee turnover as a predictor of firm performance. Journal of Management, 39(3), 573–603.

[15] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29 (pp. 3315–3323).

[16] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[17] Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. Journal of Applied Psychology, 102(3), 530–545.

[18] Huselid, M. A. (2018). The science and practice of workforce analytics: Introduction to the HRM special issue. Human Resource Management, 57(3), 679–684.

[19] IBM. (n.d.). employee-attrition-aif360 [Source code and data]. GitHub. https://github.com/IBM/employee-attrition-aif360

[20] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (pp. 43:1–43:23).

[21] Levenson, A. (2011). Using targeted analytics to improve talent decisions. People & Strategy, 34(2), 34–43.

[22] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774).

[23] Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR analytics. The International Journal of Human Resource Management, 28(1), 3–26.

[24] Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

[25] Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281.

[26] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (pp. 625–632).

[27] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[28] Prasadula, N. (2022). HR analytics: A modern tool in HR for predictive decision making. Journal of Management, 6(3), 51–63.

[29] Rasmussen, T., & Ulrich, D. (2015). Learning from practice: How HR analytics avoids being a management fad. Organizational Dynamics, 44(3), 236–242.

[30] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

[31] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 10(3), e0118432.

[32] Subhash, P. (2017). IBM HR analytics employee attrition & performance [Data set]. Kaggle. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

[33] Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61(4), 15–42.