Monitoring & Governance Plan — HR Attrition Model (v1.0)

Project: Predictive HR Analytics – Attrition

Model: Binary classifier predicting employee attrition risk

Model Owner: Vishnu Lal Raveendran Nair

Sponsor: HR VP

Validator: Independent reviewer

Last Updated: Aug 8, 2025

Contents

1) Purpose and Scope 4

2) Key Concepts 4

2.1 What is model risk? 4

2.2 What is risk tiering? 4

2.3 What is a model risk management framework? 4

2.4 What is drift? 5

3) Risk Tiering for the Attrition Model 5

3.1 Current classification and guardrails 7

3.2 Escalation triggers 7

3.3 Unacceptable risk check 7

4) Roles and Responsibilities (RACI) 7

5) Model Card (Snapshot) 8

6) Monitoring Plan 8

6.1 Prediction or population drift — PSI 8

6.2 Feature drift — CSI and Chi-square 9

6.3 Performance drift 10

6.4 Calibration drift 10

6.5 Target or base rate drift 10

6.6 Data quality and pipeline health 11

7) Variable-Level Monitoring 11

7.1 Build-time statistics to record 11

7.2 Acceptable ranges and caps 17

7.3 Missing values 18

7.4 Variable drift checks for submission 20

8) Acceptance criteria and triggers 20

9) Fairness and bias monitoring 22

10) Change management and versioning 23

11) Documentation, audit, and security 23

1) Purpose and Scope

This document defines how we govern, monitor, and manage the Attrition Risk Model from development through production. It covers risk tiering, roles and responsibilities, performance and health monitoring, drift detection, fairness checks, change management, and audit controls. Scope includes the production model, data pipelines, and dashboards used by HR BPs for decision support. The model is advisory only and may not be used for automated employment decisions.

2) Key Concepts

2.1 What is model risk?

Model risk is the risk of adverse consequences from decisions based on the incorrect selection, implementation, or use of a model. Model risk increases with greater model complexity, higher uncertainty about inputs and assumptions, broader extent of use, and larger potential impact.

2.2 What is risk tiering?

Risk tiering classifies potential harm and determines required controls. We align to EU-style categories: Unacceptable, High, Limited, Minimal.

- **Unacceptable:** manipulates behavior to circumvent free will or enables government social scoring. Banned.
- **High:** used in critical infrastructure, education and testing, essential services, law enforcement, migration or border control. Strict obligations before use.
- **Limited:** transparency obligations. Users should know they are interacting with or receiving outputs from AI. Our attrition model fits here as advisory decision support.
- **Minimal:** minimal or no risk, such as spam filters or video game AI.

2.3 What is a model risk management framework?

A structured set of policies and processes to identify, measure, monitor, and control model risk across the lifecycle: development, validation, approval, deployment, ongoing monitoring, change management, and retirement.

2.4 What is drift?

Drift means the statistical properties have changed so that training assumptions no longer hold. We monitor:

- Prediction drift: score distribution shift using PSI.
- Feature drift: input distribution shift using CSI for numeric and Chi-square for categorical.
- Performance drift: AUC and F1 decline, and calibration shift.
- Target drift: change in base rate of attrition.

3) Risk Tiering for the Attrition Model

Risk profiling framework and governance mapping

Category	**Definition**	**Examples**	**Allowed?**	**Governance for this project**
Unacceptable	Manipulates behavior to circumvent free will; enables government social scoring	Covert nudging to suppress complaints; cross-domain social scoring of employees	Banned	We will not build or deploy any such system. Screening at design review; record “UR-NA” in approvals.

High	AI used in critical infrastructures, education and testing, essential services, law enforcement, migration or border control	Automated employee triage that directly controls access to essential employment services or materially affects rights without human override	Allowed with strict obligations	If scope expands to materially influence employment decisions, escalate to High: independent validation, DPA or PIA, bias and robustness testing, human-in-the-loop, activity logs, incident playbooks, Risk and Legal approval.
Limited	Transparency obligations	Advisory attrition scores and analytics dashboards	Allowed with disclosure	**Current classification:** Limited. Require dashboard notice, use policy, monitoring and drift controls, fairness checks, and change management.
Minimal	Minimal or no risk	Spam filters, game AI	Allowed	Not applicable here.

3.1 Current classification and guardrails

Classification: **Limited Risk**. The model is advisory only and must not be used for automated or punitive HR decisions.

Required guardrails: transparency banner on dashboards; human oversight with documented rationale; scope limits that prevent adverse action workflows from scores alone; quarterly fairness checks; monitoring per §6 with documented remediation.

3.2 Escalation triggers

Escalate to **High Risk** and seek Risk and Legal approval before deployment if scores or explanations materially determine compensation, promotion, termination, scheduling, or gate access to benefits or opportunities.

3.3 Unacceptable risk check

No manipulation to circumvent free will and no government social scoring. If proposed, halt and classify as Unacceptable.

4) Roles and Responsibilities (RACI)

Activity	**Model Owner**	**Data Eng**	**HR BP**	**Validator**	**Risk/Legal**	**IT Ops**
Define business objective	R	C	A	C	C	I
Feature engineering	R	C	I	C	I	I
Model training and docs	R	C	I	C	I	I
Independent validation	I	I	I	A	C	I
Approval to deploy	A	C	C	R	C	C
Monitoring monthly	R	C	I	C	I	C
Fairness and bias review	R	I	C	C	A	I

Incident response	A	C	C	C	C	R
Change management and versioning	A	C	I	C	C	R

Legend: R Responsible, A Accountable, C Consulted, I Informed.

5) Model Card (Snapshot)

- Algorithm: Logistic Regression with class weights. Benchmarked RF, XGB, and NN.
- Training data: IBM HR dataset plus engineered features such as RoleTenureRatio, PromotionStagnation, Compa Ratio, JobHoppingCategory, PromotionFlag.
- Holdout: 30 percent stratified split.
- Primary metric: ROC AUC. Secondary: F1 (positive), Precision, Recall, Brier calibration.
- Intended use: prioritize outreach and retention conversations; not for adverse action.

Initial fit statistics

- Train AUC: **0.8978**, Test AUC: **0.8362**
- Train F1: **0.5828**, Test F1: **0.5202**
- Precision and Recall (Test): **0.3816** and **0.8169**
- Calibration (Brier): **0.1692**
- Confusion matrix at threshold 0.50: **[TP=58, FP=94, TN=276, FN=13]**

6) Monitoring Plan

Cadence: monthly with quarterly deep dive.

Artifacts: monitoring dashboard, monthly PDF or CSV report, alert log, remediation tickets.

Alert channels: email to Model Owner, Validator, HR Ops, with copy to Risk.

6.1 Prediction or population drift — PSI

- • What: compare current score distribution versus baseline.

- Method: Population Stability Index (PSI) = $\Sigma(\text{Actual\%} - \text{Expected\%}) \times \ln(\text{Actual\%} / \text{Expected\%})$, using ~10 quantile bins on the model's predicted probabilities; apply additive smoothing if any expected bin share is <1%.
- Thresholds: PSI < 0.10 no action; 0.10–0.20 report and review; ≥0.20 refit or challenge.
- Action: if PSI ≥0.20, open a change request, run a challenger, and compare AUC, F1, and Brier.

6.2 Feature drift — CSI and Chi-square

- Numeric features: CSI, same formula as PSI.

  • Categorical features (Nominal/Ordinal): Chi-Square with $\chi^2 = \Sigma (O - E)^2 / E$, where O are holdout counts and E are expected counts from build proportions scaled to the holdout total (E_i = total_holdout × $p\_build,_i$). Use p-value with df = k − 1 (k = number of categories); flag if $p < 0.05$. For df = 1, reference values: 3.841 (p=.05), 5.024 (p=.025), 6.635 (p=.01).
- Flags: CSI thresholds mirror PSI. Chi-square $p < 0.05$ is a shift.

Decision playbook for feature drift

- Measures:
  - Numeric features: CSI per feature.
  - Categorical features: Chi-Square vs build distribution.
- Thresholds and actions:
  - CSI < 0.10 or Chi-Square $p \geq 0.05$: monitor only.
  - $0.10 \leq \text{CSI} < 0.20$ or $p < 0.05$ on non-critical features: investigate the upstream data source, review recent business changes, confirm caps and floors are still appropriate, update preprocessing if needed, document.

  - CSI ≥ 0.20 or $p < 0.05$ on critical features:
    1. Validate no schema change or mapping error.
    2. Re-evaluate domain caps and outlier handling.
    3. Recalibrate model threshold if model metrics hold.
    4. If AUC or F1 drops ≥ 5 pp or drift persists 2 periods, retrain and reset baselines.
- Critical features for this model: OverTime, RoleTenureRatio, MonthlyIncome, Compa Ratio, YearsInCurrentRole.
- Always re-run fairness review after any preprocessing change or retrain.

6.3 Performance drift

Trigger when AUC drops by ≥ 5 percentage points versus baseline or F1 drops by ≥ 5 percentage points for at least two consecutive months. Action: review features or pipeline, recalibrate threshold, or refit.

6.4 Calibration drift

Metric: Brier score and reliability curve. Trigger when Brier worsens by ≥ 10 percent versus baseline. Action: recalibrate or refit.

6.5 Target or base rate drift

Metric: change in monthly attrition rate ± 3 percentage points. Action: investigate upstream changes and assess need for refit.

Decision playbook when the target rate changes

- Monitor: monthly attrition rate vs baseline.
- Thresholds and actions:
  - Within ±3 pp: monitor only.

    - Outside ±3 pp for 1 month but within ±5 pp: investigate drivers in HR operations, hiring or policy; check recent pipeline changes; log outcome.
    - ≥ ±5 pp for 1 month or ±3 pp for 2 consecutive months: recalibrate decision threshold on the latest 3 months, recheck AUC, F1 and Brier, re-evaluate fairness.
    - Persistent shift ≥ ±3 pp for 3 consecutive months: retrain with current data window, recompute class weights, re-set baseline metrics and drift references.
- Ownership: Model Owner with HR Analytics Lead. Validator signs off before threshold or retrain changes.

6.6 Data quality and pipeline health

Checks: schema, nulls, out-of-range after caps, duplicates, late arrivals, join keys, referential integrity. Action: block scoring on critical failures, alert, and remediate.

7) Variable-Level Monitoring

7.1 Build-time statistics to record

For every input feature, store:

- **Numeric:** mean, median, standard deviation, p1, p99, caps, floors, impute value.
- **Categorical:** valid categories, build distribution percent, rare bucket rule, impute value.

Inputs used by the LR pipeline at the semantic level

Engineered features

- Compa Ratio = 12 × MonthlyIncome ÷ Market Median (USD)
- PromotionStagnation = YearsSinceLastPromotion ÷ YearsAtCompany, 0 if YearsAtCompany = 0
- RoleTenureRatio = YearsInCurrentRole ÷ YearsAtCompany, 0 if YearsAtCompany = 0

- JobHoppingCategory from NumCompaniesWorked and JobHoppingIndex, then original index dropped
- PromotionFlag = “No Promotion” if YearsSinceLastPromotion equals YearsAtCompany, else “Had Promotion”

Log transforms used for skewed features

PromotionStagnation_log2, PercentSalaryHike_log2, MonthlyIncome_log2, DistanceFromHome_log2, YearsInCurrentRole_log2

Note: monitor the original variables as the primary units. Log features are diagnostic.

Numeric features to monitor

Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, PercentSalaryHike, TrainingTimesLastYear, YearsInCurrentRole, Compa Ratio, PromotionStagnation, RoleTenureRatio, plus the five log diagnostics listed above.

Categorical features to monitor

BusinessTravel, Department, EducationField, EnvironmentSatisfaction, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, OverTime, WorkLifeBalance, PromotionFlag, JobHoppingCategory.

Dropped post-tests or by policy: Gender, RelationshipSatisfaction, Education, PerformanceRating.

7.1.b Tables for this build

Numeric – record stats and set caps and floors

Feature	**Mean**	**Median**	**Std**	**P1**	**P99**	**Cap**	**Floor**	**Impute**

Age	36.9854	36	9.1939	19.28	59	59	19.28	median
Compa Ratio	0.8405	0.8522	0.2225	0.3499	1.4812	2.5	0.5	median
DistanceFromHome	9.2527	7	8.2499	1	29	29	0	median
DistanceFromHome_log2	2.8519	3	1.2554	1	4.9069	4.9069	1	median
MonthlyIncome	6517.1260	4969	4658.3370	1399.44	19605.44	19605.44	1399.44	median
MonthlyIncome_log2	12.3485	12.279	0.9511	10.4516	14.259	14.259	10.4516	median
PercentSalaryHike	15.2313	14	3.6680	11	24	100	0	median
PercentSalaryHike_log2	3.9863	3.9069	0.3102	3.585	4.6439	4.6439	3.585	median
PromotionStagnation	0.2956	0.1667	0.3445	0	1	1	0	median
PromotionStagnation_log2	0.3277	0.2224	0.3548	0	1	1	0	median
RoleTenureRatio	0.5874	0.6667	0.3313	0	1	1	0	median

TrainingTimesLastYear	2.7852	3	1.2865	0	6	20	0	median
YearsInCurrentRole	4.2089	3	3.5660	0	14.72	40	0	median
YearsInCurrentRole_log2	1.9933	2	1.1323	0	3.9739	3.9739	0	median

Categorical – valid sets and rules

Feature	**Valid Values**	**Build % (top)**	**Rare-Category Rule**	**Impute**
BusinessTravel	{Non-Travel, Travel_Frequently, Travel_Rarely}	Travel_Rarely: 70.9%, Travel_Frequently: 18.5%, Non-Travel: 10.6%	Bucket < 3%	mode
Department	{Human Resources, Research & Development, Sales}	Research & Development: 66.4%, Sales: 29.3%, Human Resources: 4.3%	Bucket < 3%	mode
EducationField	{Human Resources, Life Sciences,	Life Sciences: 40.9%, Medical:	Bucket < 3%	mode

	Marketing, Medical, Other, Technical Degree}	31.6%, Marketing: 10.4%, Technical Degree: 8.8%, Other: 6.4%, Human Resources: 1.9%		
EnvironmentSatisfaction	{High, Low, Medium, Very High}	Very High: 30.9%, High: 30.3%, Low: 20.8%, Medium: 18.0%	n/a	mode
JobHoppingCategory	{Chronic, Frequent, Fresher, Moderate, Stable}	Moderate: 34.9%, Stable: 23.1%, Frequent: 16.7%, Fresher: 14.1%, Chronic: 11.2%	Bucket < 3%	mode
JobInvolvement	{High, Low, Medium, Very High}	High: 58.7%, Medium: 24.9%, Very High: 10.3%, Low: 6.1%	n/a	mode
JobLevel	{Associate, Entry Level, Executive, Middle Management, Senior Management}	Associate: 37.1%, Entry Level: 36.1%, Middle Management: 14.9%, Senior	n/a	mode

		Management: 7.6%, Executive: 4.4%		
JobRole	{Healthcare Representative, Human Resources, Laboratory Technician, Manager, Manufacturing Director, Research Director, Research Scientist, Sales Executive, Sales Representative}	Sales Executive: 21.7%, Research Scientist: 19.7%, Laboratory Technician: 17.9%, Manufacturing Director: 10.0%, Healthcare Representative: 10.0%, Other: 20.7%	Bucket < 3%	mode
JobSatisfaction	{High, Low, Medium, Very High}	Very High: 31.9%, High: 30.6%, Low: 18.8%, Medium: 18.8%	n/a	mode
MaritalStatus	{Divorced, Married, Single}	Married: 45.8%, Single: 31.0%, Divorced: 23.2%	n/a	mode
OverTime	{No, Yes}	No: 73.6%, Yes: 26.4%	n/a	mode

PromotionFlag	{Had Promotion, No Promotion}	Had Promotion: 88.3%, No Promotion: 11.7%	n/a	mode
StockOptionLevel	{High, Low, Medium, None}	None: 41.6%, Low: 41.0%, Medium: 10.9%, High: 6.5%	n/a	mode
WorkLifeBalance	{High, Low, Medium, Very High}	High: 60.8%, Medium: 23.7%, Very High: 9.4%, Low: 6.0%	n/a	mode

7.2 Acceptable ranges and caps

Principles

- Use build p1 and p99 as default floors and caps.
- Domain caps take precedence over percentile caps.
- Values outside range are clipped to the nearest bound and flagged.

Domain caps

RoleTenureRatio 0.0 to 1.0; PromotionStagnation 0.0 to 1.0; Compa Ratio 0.5 to 2.5; DistanceFromHome minimum 0 with cap at build p99; YearsInCurrentRole 0 to 40; PercentSalaryHike 0 to 100; TrainingTimesLastYear 0 to 20.

Categorical validity

Values must be in the build valid set. Anything else maps to Other or Missing and is reviewed monthly.

Order of operations

Trim whitespace and standardize case; apply categorical mapping; clip numeric to caps; impute missing per §7.3 and create a MissingFlag.

7.3 Missing values

Imputation policy

- **Numeric:** impute with the **build median**. If there are material differences by **Department**, impute with the Department median.
- **Categorical:** impute with the **build mode**. If "Missing" is informative, keep an explicit **Missing** bucket.
- **Compa Ratio (special case):** if the MarketData join fails, impute **Compa Ratio** with the **Department-level median** and set **CompaRatio_MissingFlag = 1**. Review join failures weekly.

Flags

- Create **Feature_MissingFlag = 1** whenever a value is **imputed** or **clipped** to a cap/floor; otherwise 0.

Operational thresholds and actions

- **Missingness ≤ 1% (current period)**
  - *Non-critical feature:* continue standard impute; log in monthly report.
  - *Critical feature**:* continue standard impute; log and watch next period.
- **Missingness 1–5%**
  - *Non-critical:* investigate pipeline/source; keep imputation; add note to alert log.
  - *Critical:* investigate immediately; run sensitivity check on predictions with and without the feature; escalate to Model Owner.
- **Missingness 5–20%**
  - *Non-critical:* open ticket to Data Engineering; consider temporary domain defaults and tighten caps; review model performance and fairness this period.
  - *Critical:* open incident. If **AUC or F1 drops ≥ 5 pp** or a fairness band is breached, pause deployment changes and prepare a challenger.
- **Missingness > 20% in a period**

  - *Non-critical:* treat as a data-quality incident; backfill or delay scoring for affected rows if feasible.
  - *Critical:* **halt scoring** that depends on this feature or switch to a **last-good baseline model** after risk sign-off; backfill once data is corrected.

Critical features: OverTime, RoleTenureRatio, MonthlyIncome, Compa Ratio, YearsInCurrentRole.

Reporting

- Include per-feature missingness in the monthly monitoring report (PDF/CSV), and list any periods when thresholds were crossed along with actions taken.

Situation	**Non-critical feature**	**Critical feature***
Missingness ≤ 1 percent in the current period	Continue with standard impute. Log in monthly report.	Continue with standard impute. Log and watch next period.
1–5 percent	Investigate pipeline and source. Keep imputation. Add note to alert log.	Investigate immediately. Run sensitivity check on predictions with and without the feature. Escalate to Model Owner.
5–20 percent	Open ticket to Data Engineering. Consider temporary domain default values and tighten caps. Review model performance and fairness this period.	Open incident. If performance drops ≥ 5 pp or fairness band is breached, pause deployment changes and prepare challenger.

> 20 percent in a period	Treat as data quality incident. Backfill or delay scoring for affected rows if feasible.	Halt scoring that depends on this feature or switch to a last-good baseline model after risk sign-off. Run backfill once data is corrected.

* Critical features: OverTime, RoleTenureRatio, MonthlyIncome, Compa Ratio, YearsInCurrentRole.

- Reporting: include per-feature missingness in the monthly monitoring PDF or CSV and list any periods when thresholds were crossed with actions taken.

7.4 Variable drift checks for submission

Show CSI for RoleTenureRatio and MonthlyIncome, and Chi-square for OverTime and BusinessTravel.

Thresholds and actions: CSI < 0.10 monitor; 0.10 to 0.20 report and investigate; ≥ 0.20 or Chi-square p < 0.05 on a critical feature consider refit or preprocessing update. Two consecutive periods above threshold escalate to Model Owner and schedule remediation.

PSI/CSI are computed with ~10 quantile bins on the score or numeric feature.

8) Acceptance criteria and triggers

Monitor	**Metric**	**Window**	**Threshold**	**Trigger rule**	**Action**	**Owner**
Prediction drift	PSI on score distribution	Monthly	≥ 0.20	1 period ≥ 0.20 or 2 consecutiv	Open CR, run challenger,	Model Owner

				e periods ≥ 0.10	refit or rebuild	
Feature drift (numeric)	CSI per feature	Monthly	≥ 0.20	Any critical feature ≥ 0.20	Investigate pipeline, recalibrate or refit	Data Scientist
Feature drift (categorical)	Chi-square vs build	Monthly	$p < 0.05$	Any critical feature $p < 0.05$	Same as above	Data Scientist
Performance drift	AUC and F1	Rolling 2 months	≥ 5 pp drop vs baseline	Either metric drops ≥ 5 pp	Recalibrate threshold or retrain	Model Owner
Calibration	Brier score	Rolling 2 months	≥ 10 percent worse	Brier increases by ≥ 10 percent vs baseline	Recalibrate probabilities	ML Engineer
Target rate shift	Percent attrition	Monthly	± 3 pp	Base rate moves by ≥ 3 pp	Investigate upstream change and	HR Analytics Lead

					consider retrain	
Data quality	Nulls, ranges, schema	Daily	Any critical fail	Any required field null spike, out-of-range post-cap, or schema break	Halt scoring, fix upstream, reprocess backlog	Data Engineering
Fairness	Demographic parity (DI)	Quarterly and post-change	DI < 0.80 or > 1.25	Any monitored group outside band	Mitigate, adjust threshold if policy allows, consult HR Legal	Responsible AI Lead

Notes: “pp” means percentage points. PSI and CSI use 10 equal-width bins with minimum expected share of 1 percent per bin and additive smoothing. A critical feature is any driver in top 10 by SHAP or business priority such as OverTime, RoleTenureRatio, MonthlyIncome, Compa Ratio. Triggers fire only if subgroup size is at least 30.

9) Fairness and bias monitoring

Groups monitored: Gender, Age bands, Marital status with proxy caution, Department.

Metrics: selection rate disparity using the 80 percent rule, TPR and FPR parity deltas with absolute gap under 10 percentage points, calibration parity with group Brier within 10 percent of reference, and coverage n at least 30.

Cadence: quarterly and after any material change.

Actions: diagnose with group-conditioned SHAP and error analysis, mitigate with reweighting or constraints or policy-approved thresholds by group, re-evaluate with holdout and six-month backtest, consult HR Legal before production changes, document rationale and outcomes.

10) Change management and versioning

Champion and challenger maintained. Any change to data, features, loss, hyperparameters, or thresholds requires a change request with evidence.

CR must include: scope, motivation, dataset versions and hashes, metrics versus baseline, backtest, canary plan, rollback plan, and sign-offs.

Versioning: semantic versioning MAJOR.MINOR.PATCH; tag code commit, data snapshot hash, model artifact ID, feature schema, SHAP report; record training environment and dependency lockfile.

Promotion gates: pass acceptance thresholds for two consecutive evaluation windows, no high-severity incidents open, fairness within band.

Rollback: keep N-1 artifact hot, roll back within one hour via configuration, recompute impacted scores, notify consumers.

Retirement: document rationale and archive artifacts and monitoring logs per retention policy.

11) Documentation, audit, and security

Required artifacts: model card, data dictionary, EDA summary, validation report, SHAP and key driver analysis, monitoring dashboards, drift logs, fairness reports, incidents and resolutions, approvals and CRs.

Access and security: RBAC with least privilege, MFA for production changes, secrets in a vault, encryption at rest and in transit, endpoint rate limits and audit trails.

Privacy: pseudonymize employee identifiers, store only required fields, follow retention and deletion schedules, no free-text PII in features, Compa Ratio uses aggregates only.

Auditability: each score links to model version, feature vector hash, input timestamp, and decision threshold; retain logs for at least 12 months.