Back to Document Index

Model Monitoring & Performance Standards

Business Analysis Template for AI Projects

Template ID:	6.7
Category:	AI Governance, Risk & Compliance
Version:	1.0
Last Updated:	January 2026
Prerequisites:	High-Level AI Solution Architecture (1.3) AI Risk Assessment Report (6.1) Model Governance Framework (6.2)

1. Document Purpose

These Model Monitoring & Performance Standards establish a comprehensive framework for monitoring AI/ML models in production to ensure they maintain expected performance, operate safely, and deliver business value over time. Model monitoring is not optional—it is essential risk management practice and regulatory expectation for AI systems.

Key purposes of these standards include:

- Define what must be monitored for AI models in production
- Establish performance metrics, thresholds, and acceptable ranges
- Detect model performance degradation early before business impact
- Identify data drift, model drift, and concept drift
- Ensure models continue to meet fairness and bias requirements
- Monitor operational health (latency, errors, availability)
- Provide early warning system through alerts and escalations
- Enable proactive intervention and model refresh decisions
- Support regulatory compliance and audit requirements
- Track business value and ROI of AI systems
- Create accountability for model performance
- Facilitate continuous improvement of AI systems

Back to Document Index

Critical Principle: Models are not "deploy and forget." Model performance degrades over time due to data drift, changing business conditions, evolving user behavior, and other factors. Without monitoring, organizations deploy models that silently degrade, making increasingly poor predictions while appearing to function normally. Monitoring is essential to detect issues before they cause business harm.

2. When to Apply These Standards

These monitoring standards apply to all AI/ML models deployed in production, with monitoring rigor scaled based on model risk tier.

Mandatory Monitoring Scenarios:

All Production Models

Every AI/ML model serving production traffic must have monitoring. No exceptions. Even low-risk models need basic monitoring to detect unexpected issues.

High-Risk and Critical Models

Models with high or critical risk tier (Template 6.1) require comprehensive monitoring with stringent thresholds, frequent reviews, and rapid response procedures. These include models making consequential decisions about individuals, high-value business processes, or safety-critical applications.

Models Making Automated Decisions

Models making fully or partially automated decisions require monitoring of decision quality, fairness across groups, and business outcomes to ensure decisions remain sound.

Models Processing Personal Data

Models processing personal data require monitoring for privacy compliance, data minimization, and bias/fairness to protect individual rights.

Regulated Models

Models subject to regulatory oversight (financial services SR 11-7, healthcare, fair lending, etc.) must meet regulatory monitoring expectations, which often exceed standard practices.

Models in Dynamic Environments

Models operating in rapidly changing environments (fraud detection, recommendation systems, pricing, etc.) require more frequent monitoring due to higher drift risk.

Models Post-Incident

Models that have experienced performance issues, incidents, or audit findings require enhanced monitoring until confidence is restored.

Back to Document Index

Monitoring Frequency by Risk Tier:

Monitoring rigor scales with model risk:

Critical Risk Models:

- Real-time automated monitoring
- Daily manual review of key metrics
- Weekly detailed performance reports
- Monthly performance deep-dives with governance committee
- Quarterly comprehensive validation

High Risk Models:

- Automated monitoring with daily alerts
- Weekly performance review
- Monthly performance reports to governance
- Quarterly performance reviews
- Annual comprehensive validation

Medium Risk Models:

- Automated monitoring with weekly alerts
- Bi-weekly or monthly performance review
- Quarterly performance reports
- Semi-annual performance reviews
- Bi-annual validation

Low Risk Models:

- Automated monitoring with monthly alerts
- Monthly or quarterly performance review
- Semi-annual performance reports
- Annual performance review
- Annual validation (if needed)

Back to Document Index

3. Model Monitoring Framework

Comprehensive model monitoring covers five dimensions. All five dimensions must be monitored, though specific metrics vary by use case.

3.1 Model Performance Monitoring

Monitoring how well the model performs its core task—making accurate predictions or classifications.

Statistical Performance Metrics:

For Classification Models:
☐ Accuracy: Overall percentage of correct predictions
☐ Precision: Of predicted positives, how many are actually positive (minimizes false positives)
☐ Recall (Sensitivity): Of actual positives, how many are predicted positive (minimizes false negatives)
☐ F1 Score: Harmonic mean of precision and recall
☐ AUC-ROC: Area under receiver operating characteristic curve
☐ Specificity: Of actual negatives, how many are predicted negative
☐ False Positive Rate (FPR)
☐ False Negative Rate (FNR)
☐ Confusion Matrix: Detailed breakdown of predictions vs. actuals

For Regression Models:
☐ Mean Absolute Error (MAE): Average absolute difference between predicted and actual
☐ Mean Squared Error (MSE): Average squared difference (penalizes large errors more)
☐ Root Mean Squared Error (RMSE): Square root of MSE (same units as target)
☐ R-squared (R^2): Proportion of variance explained by model
☐ Mean Absolute Percentage Error (MAPE): Average percentage error

For Ranking/Recommendation Models:
☐ Precision@K: Precision for top K recommendations
☐ Recall@K: Recall for top K recommendations
☐ Mean Average Precision (MAP)
☐ Normalized Discounted Cumulative Gain (NDCG)
☐ Hit Rate@K: Percentage of times relevant item appears in top K

Business Performance Metrics:
Track how model performance translates to business outcomes:

☐ Conversion rate (if model drives conversions)

Back to Document Index

☐ Revenue impact (if model affects revenue)
☐ Cost savings (if model reduces costs)
☐ Customer satisfaction (if model impacts customers)
☐ Fraud losses prevented (for fraud models)
☐ Default rate (for credit models)
☐ Click-through rate (for recommendation models)
☐ Time to resolution (for support models)
☐ Other business-specific KPIs aligned with model purpose

Fairness Metrics (Critical for High-Risk Models):
Monitor performance across demographic groups to ensure fairness:

☐ Accuracy by group (race, gender, age, etc.)
☐ Precision by group
☐ Recall by group
☐ False positive rate by group
☐ False negative rate by group
☐ Demographic parity (equal positive prediction rates)
☐ Equal opportunity (equal recall across groups)
☐ Disparate impact ratio (should be ≥80% per EEOC guidance)

Why These Metrics Matter:
- Detect if model accuracy is degrading over time
- Identify if model is making different types of errors than during development
- Ensure model continues to meet business requirements
- Verify fairness is maintained across demographic groups
- Provide evidence for audits and regulatory reviews

Monitoring Implementation:
- Calculate metrics on holdout test set weekly or monthly
- Calculate metrics on recent production data
- Compare current metrics to baseline (development performance)
- Track trends over time
- Alert when metrics fall below thresholds
- Segment analysis by relevant dimensions (time period, geography, product, demographic)

3.2 Data Drift Monitoring

Monitoring changes in input data distribution. When data changes, model performance often degrades even if the model itself is unchanged.

Back to Document Index

What is Data Drift:
Data drift occurs when statistical properties of input features change over time. For example:

- Customer demographics shift
- Product mix changes
- Seasonal patterns evolve
- Economic conditions change
- User behavior changes
- Data collection process changes

Why Data Drift Matters:
Models trained on historical data assume future data will follow the same distribution. When data distribution changes, model
performance degrades because the model encounters scenarios it wasn't trained for.

Data Drift Detection Methods:

Univariate Drift Detection (per feature):
☐ Population Stability Index (PSI): Measures distribution change

- PSI < 0.1: No significant drift
- 0.1 ≤ PSI < 0.25: Moderate drift, investigate
- PSI ≥ 0.25: Significant drift, action required

☐ Kullback-Leibler (KL) Divergence: Measures distribution difference
☐ Jensen-Shannon Divergence: Symmetric version of KL divergence
☐ Kolmogorov-Smirnov Test: Statistical test for distribution equality
☐ Chi-Squared Test: For categorical features

Summary Statistics:
☐ Mean shift detection
☐ Standard deviation changes
☐ Min/max range changes
☐ Percentile shifts (25th, 50th, 75th)
☐ Missing value rate changes
☐ Unique value count changes (for categorical)

Multivariate Drift Detection:
☐ Adversarial validation: Train classifier to distinguish training vs. production data (high accuracy indicates drift)
☐ Maximum Mean Discrepancy (MMD): Multivariate distribution distance
☐ Principal Component Analysis (PCA): Monitor principal components

Back to Document Index

Data Quality Monitoring:
☐ Missing value rate per feature
☐ Out-of-range values
☐ Invalid values (violating business rules)
☐ Data type violations
☐ Unexpected nulls
☐ Duplicate records

Monitoring Implementation:
- Monitor all features used by model
- Calculate drift metrics daily or weekly
- Compare recent data (e.g., last 7 days) to training/baseline distribution
- Alert when drift exceeds thresholds
- Prioritize monitoring features most important to model (via feature importance)
- Track data quality metrics

3.3 Model Drift and Concept Drift Monitoring

Monitoring changes in relationship between inputs and outputs, and changes in model predictions themselves.

Model Drift (Prediction Drift):
Changes in model's predictions over time, even with similar inputs.

What to Monitor:
☐ Prediction distribution: Are predictions shifting (e.g., predicting more/fewer positives)?
☐ Prediction confidence/probability distribution: Is model becoming more/less confident?
☐ Average predicted probability
☐ Percentage of high-confidence predictions
☐ Percentage of predictions in each class (for classification)
☐ Distribution of continuous predictions (for regression)

Concept Drift:
The relationship between features and target variable changes. For example:
- What made a customer likely to churn last year may differ from this year
- Economic conditions change relationship between variables
- Competitive landscape shifts customer behavior
- Regulations change relationship between inputs and outcomes

Why Concept Drift Matters:
Even if data distribution stays similar, the underlying patterns may change. Model trained on

Back to Document Index

old patterns makes poor
predictions on new patterns.

Concept Drift Detection:
☐ Monitor actual outcomes vs. predictions: Are predictions becoming less accurate over time?
☐ Residual analysis: Are prediction errors increasing or showing new patterns?
☐ Feature importance changes: Are different features becoming more/less important?
☐ Model recalibration: Does model need different decision thresholds?
☐ A/B testing: Does retrained model significantly outperform current model?

Monitoring Implementation:

- Track prediction distribution daily or weekly
- Compare current prediction distribution to baseline
- Monitor prediction confidence trends
- Calculate performance on recent labeled data
- Alert when predictions shift significantly
- Investigate whether shift is due to data drift or concept drift

3.4 Operational Health Monitoring

Monitoring system-level performance to ensure the model is available, responsive, and functioning correctly.

System Performance Metrics:

Latency and Response Time:
☐ Average prediction latency (time from request to response)
☐ P50, P95, P99 latency (percentile latencies)
☐ Batch processing time (if applicable)
☐ End-to-end response time including pre/post processing

Thresholds:

- P95 latency should not exceed [X] milliseconds (define based on SLA)
- Average latency should remain below [X] milliseconds

Availability and Uptime:
☐ Model availability (percentage of time model is accessible)
☐ Uptime percentage
☐ Downtime incidents and duration
☐ Mean Time Between Failures (MTBF)

Back to Document Index

☐ Mean Time To Recovery (MTTR)

Thresholds:
- Availability target: 99.9% (or higher for critical models)
- Maximum acceptable downtime per incident
- Maximum acceptable downtime per month

Error Rates:
☐ Prediction error rate (technical errors, not accuracy errors)
☐ Timeout rate
☐ Failed requests
☐ Invalid input rate (requests model can't process)
☐ Exception/crash rate

Thresholds:
- Technical error rate should be <0.1%
- Invalid input rate tracked but may not be model's fault

Throughput and Capacity:
☐ Requests per second (RPS)
☐ Predictions per minute/hour/day
☐ Batch size (if batch processing)
☐ Queue depth (if using queues)
☐ Resource utilization (CPU, memory, GPU)

Monitoring:
- Ensure throughput meets demand
- Alert if capacity constraints approached
- Plan for capacity scaling

Data Completeness:
☐ Percentage of requests with all required features
☐ Feature null rate
☐ Feature availability
☐ Data pipeline health

Model Versioning:
☐ Current model version in production
☐ Model deployment date
☐ Model rollback capability
☐ A/B testing assignments (if running experiments)

Dependencies:

Back to Document Index

☐ Health of upstream data sources
☐ Health of downstream consumers
☐ Third-party API health (if model depends on external data)
☐ Database health

Why Operational Monitoring Matters:
- Technical failures can be mistaken for model performance issues
- Latency impacts user experience
- Downtime causes business disruption
- Capacity constraints lead to degraded performance
- Operational issues must be distinguished from model quality issues

Monitoring Implementation:
- Real-time operational monitoring dashboards
- Infrastructure monitoring tools (Prometheus, Grafana, Datadog, New Relic, etc.)
- Alert on SLA violations
- Track operational metrics 24/7
- Page on-call engineer for critical operational issues

3.5 Business Value and ROI Monitoring

Monitoring whether the model continues to deliver expected business value and return on investment.

Business Impact Metrics:
Track metrics aligned with original business case (Template 1.1):

Revenue Metrics:
☐ Revenue attributed to model
☐ Revenue lift vs. baseline
☐ Average transaction value influenced by model
☐ Customer lifetime value impact

Cost Metrics:
☐ Cost savings from automation
☐ Operational efficiency gains
☐ Reduced manual review time
☐ Fraud losses prevented
☐ Default/risk costs avoided

Customer Metrics:

☐ Customer satisfaction (CSAT) for model-influenced interactions
☐ Net Promoter Score (NPS) impact
☐ Customer retention influenced by model
☐ Customer complaints related to model
☐ Conversion rate

Operational Metrics:
☐ Processing time reduction
☐ Decision speed improvement
☐ Resource utilization efficiency
☐ Quality improvement
☐ Error reduction

Adoption and Usage Metrics:
☐ Model utilization rate (% of eligible decisions using model)
☐ User adoption by stakeholder group
☐ Overrides (when humans override model recommendations)
☐ Override reasons and patterns
☐ User satisfaction with model

Return on Investment (ROI):
☐ Actual ROI vs. projected ROI from business case
☐ Payback period
☐ Total Cost of Ownership (TCO): Development + operations + maintenance
☐ Cost per prediction
☐ Value per prediction

Comparative Analysis:
☐ Model performance vs. baseline (previous approach)
☐ Model performance vs. champion/challenger models
☐ A/B test results (if running experiments)
☐ Human performance vs. model performance (where comparable)

Why Business Monitoring Matters:

- Technical performance doesn't guarantee business value
- Business conditions change—value delivered may shift
- ROI must be sustained to justify ongoing investment
- Stakeholder satisfaction affects model adoption
- Business metrics inform model refresh priorities

Monitoring Implementation:

- Define business metrics aligned with business case
- Track business metrics monthly or quarterly

Back to Document Index

- Compare actuals to projections
- Report business value to stakeholders and governance committees
- Justify continued investment through demonstrated value

Back to Document Index

4. Alert Configuration and Escalation

Effective monitoring requires actionable alerts that route issues to appropriate teams for timely response.

4.1 Alert Severity Levels

Level 1: Critical (Immediate Response Required)

Trigger Conditions:
- Model unavailable or experiencing severe technical failures (>5% error rate)
- Performance degradation >20% from baseline on critical metrics
- Severe fairness violations (disparate impact <60%)
- Data quality catastrophic (>30% missing values on critical features)
- Severe concept drift detected
- Safety incident or near-miss
- Regulatory threshold breach

Response:
- Immediate investigation by on-call team
- Page model owner and technical lead
- Assess whether to pause model and revert to fallback
- Convene incident response team if needed
- Notify governance committee within 24 hours
- Document incident for post-mortem

Response Time: <1 hour investigation start, <4 hours initial containment

Level 2: High (Urgent, Same-Day Response)

Trigger Conditions:
- Performance degradation 10-20% from baseline
- Moderate data drift (PSI 0.25-0.50)
- Fairness concerns (disparate impact 60-80%)
- Latency exceeding SLA but still functional
- Elevated error rate (1-5%)
- Business metric declining significantly

Response:
- Investigation by model team within business hours
- Assess need for model refresh or retraining
- Notify model owner and stakeholders

Back to Document Index

- Create remediation ticket with priority
- Update governance at next regular meeting

Response Time: <4 hours investigation start, <24 hours action plan

Level 3: Medium (Important, Multi-Day Response)

Trigger Conditions:

- Performance degradation 5-10% from baseline
- Minor data drift (PSI 0.10-0.25)
- Fairness metrics trending unfavorable but still compliant
- Operational metrics approaching thresholds
- Business metrics below target but not critical

Response:

- Investigation within 2-3 business days
- Root cause analysis
- Remediation plan with timeline
- Document in monitoring log
- Include in regular performance review

Response Time: <3 days investigation, <2 weeks remediation plan

Level 4: Low (Informational, Monitoring)

Trigger Conditions:

- Performance degradation <5% from baseline
- Minor data drift (PSI <0.10)
- Operational metrics within normal variance
- Business metrics trending but not concerning
- Informational events (model updates, configuration changes)

Response:

- Log for trending analysis
- Include in regular reports
- No immediate action required
- Review during periodic performance reviews

Response Time: Addressed in regular monitoring cycle

Back to Document Index

4.2 Escalation Procedures

Escalation Path:

Level 1: Model Operations Team
- First line of response for all alerts
- Technical troubleshooting
- Initial assessment and containment
- Escalate based on severity and complexity

Level 2: Model Owner + Technical Lead
- Engaged for High and Critical alerts
- Decision authority for model interventions
- Coordinate with stakeholders
- Approve model changes or rollbacks

Level 3: Model Risk Manager / Governance Committee
- Engaged for Critical alerts and significant issues
- Governance and risk perspective
- Cross-functional coordination
- Escalation to executive leadership if needed

Level 4: Executive Leadership / Board
- Engaged for incidents with significant business/regulatory/reputational impact
- Strategic decisions
- External communications
- Resource allocation for major remediation

Escalation Timing:
- Critical alerts: Immediate escalation to Level 2, Level 3 within 4 hours
- High alerts: Escalation to Level 2 within 4 hours, Level 3 if unresolved in 24 hours
- Medium alerts: Escalation to Level 2 within 3 days if unresolved
- Low alerts: Escalation only if pattern emerges

Escalation Criteria:
- Severity of impact
- Duration of issue
- Inability to resolve at current level
- Regulatory implications
- Reputational risk
- Pattern of recurring issues

Back to Document Index

4.3 Alert Configuration Best Practices

Set Appropriate Thresholds

Thresholds should be:

- Based on baseline performance during development/validation
- Informed by business tolerance for degradation
- Statistically significant (avoid alerting on noise)
- Reviewed and adjusted based on alert history
- More stringent for high-risk models

Avoid Alert Fatigue

Too many low-value alerts cause teams to ignore all alerts:

- Start with conservative thresholds, tighten over time
- Prioritize alerts by severity
- Aggregate related alerts (don't send 100 separate alerts for same issue)
- Provide actionable information in alerts
- Review and disable noisy alerts

Make Alerts Actionable

Every alert should clearly communicate:

- What triggered the alert
- Current value vs. threshold
- Trend (is it getting worse?)
- Which model is affected
- Links to dashboards and runbooks
- Suggested actions
- Escalation path

Use Multiple Communication Channels

Route alerts appropriately:

- Critical: Page/SMS + Email + Dashboard + Incident management system
- High: Email + Slack/Teams + Dashboard
- Medium: Email + Dashboard
- Low: Dashboard only
- Don't spam everyone—route to responsible teams

Document Alert Response

For every alert:

- Log alert occurrence
- Document investigation and findings
- Record actions taken
- Update runbooks if new response developed
- Track time to detection, investigation, resolution
- Use for continuous improvement

Back to Document Index

Back to Document Index

5. Monitoring Dashboards and Reporting

Dashboards and reports make monitoring data accessible, actionable, and support decision-making at different organizational levels.

5.1 Real-Time Operational Dashboard

For model operations teams monitoring day-to-day performance.

Purpose: Enable rapid detection and response to operational and performance issues

Audience: Model operations team, SRE/DevOps, on-call engineers

Refresh Frequency: Real-time or every 1-5 minutes

Key Metrics to Display:

System Health (Top Section):
- Overall model status: ✅ Healthy | ⚠️ Degraded | ❌ Down
- Current availability percentage
- Active alerts by severity
- Current request rate (RPS)
- Average latency (last 1 hour)
- Error rate (last 1 hour)

Performance Metrics (Middle Section):
- Primary model metric (e.g., accuracy, precision, recall) - current value vs. threshold
- Prediction distribution (last 24 hours)
- Confidence/probability distribution
- Fairness metrics status (if applicable)

Data Health (Middle Section):
- Feature availability
- Missing value rates (top 5 features)
- Data quality alerts
- Data drift indicators

Time Series Trends (Bottom Section):
- Latency over time (last 24 hours, 7 days)
- Request volume over time
- Error rate over time

- Primary performance metric over time
- Prediction distribution over time

Alerts Panel:
- Active alerts (last 24 hours)
- Alert severity distribution
- Recently resolved alerts
- Alert acknowledgment status

Quick Actions:
- Links to detailed investigation dashboards
- Links to runbooks
- Model rollback capability (if permissions allow)
- Alert acknowledgment buttons

5.2 Model Performance Dashboard

For model owners, data scientists, and stakeholders reviewing model performance.

Purpose: Understand model performance trends, degradation patterns, and health

Audience: Model owners, data scientists, ML engineers, product managers

Refresh Frequency: Daily or hourly

Key Sections:

Executive Summary:
- Model name and current version
- Overall health score (composite metric)
- Days since last deployment
- Days since last performance review
- Key performance indicators vs. targets
- Active issues requiring attention

Statistical Performance:
- All relevant metrics for model type (accuracy, precision, recall, etc.)
- Current value | Baseline (development) | Threshold | Status
- Trend indicators (↑ improving, → stable, ↓ degrading)
- Time series charts showing metric evolution
- Performance by segment (if applicable—geography, product, channel, etc.)

Back to Document Index

Fairness Analysis:
- Performance metrics by demographic group (table)
- Disparate impact ratios
- Fairness metric trends
- Visual comparison across groups

Data Drift Analysis:
- PSI scores for top 10 most important features
- Data drift heatmap (all features)
- Distributions: Training vs. Current (for drifted features)
- Feature importance changes

Prediction Analysis:
- Prediction distribution histogram
- Prediction trends over time
- Confidence distribution
- Prediction vs. actual (if ground truth available)
- Residual analysis (for regression)

Business Impact:
- Business metrics aligned with business case
- ROI calculation
- Value delivered vs. target
- Adoption and usage metrics

5.3 Executive Dashboard

For executives and governance committees overseeing AI/ML portfolio.

Purpose: Portfolio-level view of all AI/ML models, risks, and value

Audience: Executive leadership, AI Governance Committee, Model Risk Committee, Board

Refresh Frequency: Weekly or monthly

Key Sections:

Portfolio Health:
- Total number of production models
- Models by risk tier (Critical, High, Medium, Low)

Back to Document Index

- Models by health status (Healthy, Degraded, At Risk, Down)
- Models requiring attention (count)
- Models pending validation/review (count)

Risk and Compliance:
- High/Critical models with performance issues
- Models with open audit findings
- Models with fairness concerns
- Models overdue for validation
- Models with active incidents
- Regulatory compliance status

Business Value:
- Total value delivered by AI portfolio ($ or other metric)
- ROI by model
- Top performing models by business impact
- Underperforming models
- New models deployed (last quarter)
- Models retired (last quarter)

Trends:
- Portfolio health trend (last 6-12 months)
- Incident frequency trend
- Value delivery trend
- Model development pipeline status

Alerts and Escalations:
- Critical/High alerts in last period
- Escalations requiring governance attention
- Pending decisions

5.4 Regular Reporting

Daily Reports (for Critical/High Risk Models):
- Automated report to model operations team
- Key metrics vs. thresholds
- Alerts generated in last 24 hours
- Actions taken

Weekly Reports (for High/Medium Risk Models):
- Performance summary
- Week-over-week trends

Back to Document Index

- Data drift summary
- Issues requiring attention
- Distributed to model owner and stakeholders

Monthly Reports (for All Models):

- Comprehensive performance review
- Business value delivered
- Comparison to baseline and targets
- Data and model drift analysis
- Fairness assessment
- Operational health summary
- Issues and remediation status
- Distributed to model owner, governance, stakeholders

Quarterly Reports (for All Models):

- Deep-dive performance analysis
- Validation results (if quarterly validation conducted)
- ROI and business value analysis
- Recommendations (continue, retrain, retire)
- Presented to Model Risk Committee or AI Governance Committee
- Formal documentation for audit trail

Ad Hoc Reports:

- Incident reports
- Post-mortem analysis
- Response to stakeholder inquiries
- Regulatory examinations
- Audit requests

Back to Document Index

6. Performance Review Process

Regular performance reviews ensure sustained model health and inform decisions about model refresh, retraining, or retirement.

6.1 Periodic Performance Review

Purpose:
- Comprehensive assessment of model performance
- Determine if model continues to meet requirements
- Decide on model refresh, retraining, or retirement
- Identify improvements for next model version
- Update risk assessment if needed

Frequency by Risk Tier:
- Critical: Monthly
- High: Quarterly
- Medium: Semi-annually
- Low: Annually

Review Participants:
- Model owner (chair)
- Data scientist/ML engineer
- Model validator (for high/critical)
- Business stakeholder
- Compliance/risk representative (for high/critical)

Review Agenda:

1. Model Overview and Context
   - Model purpose and use case
   - Current model version and deployment date
   - Changes since last review
   - Previous review recommendations status

2. Performance Assessment
   - Statistical performance vs. baseline and thresholds
   - Trends (improving, stable, degrading)
   - Performance by segment
   - Fairness metrics
   - Root cause analysis for any degradation

Back to Document Index

3. Data and Model Drift
   - Data drift assessment (PSI scores, feature analysis)
   - Model drift assessment (prediction distribution)
   - Concept drift indicators
   - Significance and business impact of drift

4. Operational Health
   - Availability and uptime
   - Latency and throughput
   - Error rates
   - Capacity and scaling
   - Incidents and issues

5. Business Value
   - Business metrics vs. targets
   - ROI assessment
   - Stakeholder feedback
   - Adoption and usage
   - Comparison to baseline/alternatives

6. Compliance and Risk
   - Fairness and bias status
   - Privacy compliance
   - Regulatory compliance
   - Audit findings status
   - Risk assessment updates needed

7. Recommendations and Actions
   - Continue as-is
   - Retrain with recent data
   - Refresh (rebuild with new features/algorithms)
   - Retire and replace
   - Specific improvements needed
   - Timeline for actions

8. Documentation
   - Meeting minutes
   - Decisions and rationale
   - Action items with owners and due dates
   - Updates to model documentation

Back to Document Index

6.2 Retraining Decision Framework

Criteria for deciding when to retrain models:

Retrain Immediately If:

- Performance degradation >20% from baseline on critical metrics
- Severe data or concept drift (PSI >0.5)
- Fairness violations (disparate impact <60%)
- Business metric significantly below target
- Regulatory threshold breach
- Major data collection or business process change

Retrain Soon (Within 1-3 Months) If:

- Performance degradation 10-20% from baseline
- Moderate drift (PSI 0.25-0.50)
- Fairness concerns (disparate impact 60-80%)
- Business metric trending unfavorable
- Scheduled periodic retraining due

Retrain on Schedule If:

- Performance stable but regular retraining planned (e.g., quarterly for high-risk models)
- Preventive retraining to avoid degradation
- New data available that may improve performance

Do Not Retrain If:

- Performance remains strong
- Minimal drift
- Retraining unlikely to improve performance
- Cost/effort exceeds benefit

Retraining Process:

1. Collect recent labeled data
2. Assess data quality and representativeness
3. Retrain model using same or updated methodology
4. Validate retrained model
5. A/B test if possible (champion vs. challenger)
6. Conduct bias/fairness testing
7. Obtain approval based on risk tier
8. Deploy with monitoring
9. Document retraining rationale and results

Back to Document Index

Back to Document Index

7. Responding to Performance Degradation

Systematic approach to investigating and responding to model performance issues.

7.1 Investigation Process

Step 1: Detect and Triage

- Alert triggers indicating performance degradation
- Assess severity and business impact
- Determine urgency of response
- Assign investigation owner

Step 2: Initial Assessment

- Confirm degradation is real (not data anomaly or monitoring issue)
- Determine scope (all predictions or specific segments)
- Check for obvious causes:
  - Recent model changes or deployments
  - Data pipeline failures
  - Operational issues (latency, errors)
  - Known incidents or outages
  - Upstream system changes

Step 3: Deep Dive Analysis

Performance Analysis:

- Which metrics are degraded?
- By how much?
- Since when?
- Sudden drop or gradual degradation?
- Affecting all segments or specific ones?

Data Analysis:

- Has data distribution changed? (drift analysis)
- Data quality issues? (missing values, invalid values, outliers)
- Feature availability changes?
- New data sources or pipeline changes?

Model Behavior Analysis:

- Are predictions shifting?
- Is confidence changing?
- Are different types of errors occurring?
- Feature importance shifts?

External Factors:
- Business changes (new products, markets, processes)
- Seasonal effects
- Economic conditions
- Competitive changes
- Regulatory changes

Step 4: Root Cause Identification
Determine primary cause:
- Data drift (input distribution changed)
- Concept drift (relationship between X and Y changed)
- Data quality degradation
- Model staleness (trained on old patterns)
- Technical/operational issue
- External/business environment change
- Combination of factors

Step 5: Develop Action Plan
Based on root cause, determine response:
- Quick fixes (if available)
- Retraining schedule
- Data quality remediation
- Feature engineering updates
- Model refresh (rebuild with new approach)
- Fallback to previous model version
- Manual review processes as interim measure

Step 6: Implement and Monitor
- Execute action plan
- Monitor closely to ensure effectiveness
- Document investigation and actions
- Update runbooks and documentation
- Prevent recurrence through process improvements

7.2 Intervention Options

Immediate Interventions (Hours to Days):

Pause Model:
- Stop model predictions and revert to fallback
- Use when: Model causing harm or severe degradation

Back to Document Index

- Risk: Business disruption if no adequate fallback

Rollback to Previous Version:

- Deploy previous known-good model version
- Use when: Recent deployment caused degradation
- Risk: Previous version may also degrade over time

Adjust Decision Thresholds:

- Change classification thresholds without retraining
- Use when: Calibration issue but model still useful
- Risk: May not address underlying issue

Increase Human Review:

- Route more predictions to human review
- Use when: Model unreliable but human judgment available
- Risk: Increased cost and reduced automation benefit

Short-Term Interventions (Weeks):

Retrain with Recent Data:

- Quick retraining with fresh data
- Use when: Data/concept drift but methodology still sound
- Risk: May not improve if underlying methodology inadequate

Data Quality Remediation:

- Fix data pipeline issues
- Use when: Data quality degradation is root cause
- Risk: Time to identify and fix all data issues

Feature Engineering Updates:

- Add new features or remove degraded features
- Use when: Feature importance shifted
- Risk: Requires validation and testing

Long-Term Interventions (Months):

Model Refresh:

- Rebuild model with new algorithms, features, data
- Use when: Fundamental rethinking needed
- Risk: Time and resource intensive

Architecture Changes:

- Change model architecture or ensemble approach

Back to Document Index

- Use when: Current approach fundamentally limited
- Risk: Significant development and validation effort

Retire Model:

- Decommission model entirely
- Use when: Model no longer viable or valuable
- Risk: Loss of automation, revert to manual processes

Back to Document Index

8. Best Practices for Model Monitoring

1. Monitor from Day One

Implement monitoring before deployment, not after. Baseline metrics during validation become monitoring thresholds. Having monitoring infrastructure ready at launch enables rapid issue detection.

2. Establish Meaningful Baselines

Monitoring requires a baseline for comparison. Use validation/test set performance as baseline. Update baselines when retraining. Without a clear baseline, cannot determine if performance is degrading.

3. Monitor What Matters to Business

Balance technical metrics (accuracy) with business metrics (revenue, customer satisfaction). Technical performance doesn't guarantee business value. Monitor metrics stakeholders care about.

4. Scale Monitoring to Risk

Critical models need comprehensive, frequent monitoring. Low-risk models need lighter monitoring. Don't apply one-size-fits-all—it wastes resources or creates gaps. Risk tier (Template 6.1) should drive monitoring rigor.

5. Automate Everything Possible

Manual monitoring doesn't scale and is unreliable. Automate metric calculation, dashboard updates, alerting, reporting. Humans review dashboards and investigate alerts, but automation does the heavy lifting.

6. Make Monitoring Observable

Dashboards should be accessible to all stakeholders. Don't lock monitoring in technical systems only data scientists can access. Transparency builds trust and enables collaborative problem-solving.

7. Alert on Actionable Issues

Every alert should have clear action. If alert doesn't drive action, it's noise. Review alert history regularly and eliminate alerts that are ignored or lead nowhere.

8. Distinguish Correlation from Causation

When performance degrades, many things may have changed. Investigate thoroughly to identify the root cause, not just correlated events. Data drift may coincide with performance drop but may not cause it.

Back to Document Index

9. Monitor Fairness Continuously

Bias can emerge post-deployment even if the model was fair initially. Continuously monitor fairness metrics across demographic groups. Don't assume deployment-time fairness is sustained.

10. Prepare for Drift

All models drift eventually. Plan for retraining from the start. Have labeled data pipeline for ongoing model validation. Budget for periodic retraining. Drift is not failure—it's the expected reality of ML in production.

11. Document Everything

Document monitoring configuration, threshold rationale, alert responses, investigation findings, remediation actions. This creates institutional knowledge and an audit trail. In the future you will thank the current you for documentation.

12. Test Monitoring and Alerting

Periodically test that alerts fire correctly, dashboards display accurately, escalations route properly. Don't assume monitoring works—verify it. Run drills where you inject synthetic degradation.

Back to Document Index

9. Common Pitfalls in Model Monitoring

1. Deploy and Forget

Deploying model without monitoring, assuming it will work indefinitely. Reality: All models degrade. Without monitoring, degradation is invisible until a major incident occurs. Never deploy without a monitoring plan.

2. Monitoring Only Technical Metrics

Tracking accuracy but ignoring business outcomes. The model may maintain accuracy while business value evaporates due to changing conditions. Monitor what matters to stakeholders, not just what's easy to measure.

3. Setting Arbitrary Thresholds

Alert thresholds not based on actual baseline, business tolerance, or statistical significance. Results in either alert fatigue (too many false alarms) or missed issues (thresholds too loose). Use baseline performance and business requirements to set thresholds.

4. Alert Fatigue

Too many low-value alerts causing teams to ignore all alerts, including critical ones. Over-alerting is as dangerous as under-alerting. Tune alert thresholds, aggregate related alerts, prioritize by severity, review and disable noisy alerts.

5. Monitoring Without Action

Tracking metrics but not acting on degradation. Monitoring value comes from driving intervention, not just observing decline. If not prepared to act on alerts, monitoring is wasted effort.

6. Ignoring Data Drift

Monitoring model performance but not input data distribution. Performance may seem fine until it suddenly collapses because data has shifted outside the training distribution. Data drift is a leading indicator—catch it early.

7. No Ground Truth for Validation

Deploying model without plan to obtain actual outcomes (labels) for ongoing validation. Can monitor predictions but can't validate accuracy. Critical to have feedback loop capturing ground truth.

8. Monitoring in Silos

Model team monitors technical metrics, business team monitors business metrics, no coordination. Results in fragmented view and delayed issue detection. Integrate monitoring across teams.

Back to Document Index

9. Stale Baselines

Comparing current performance to outdated baseline from initial development years ago. After retraining, update the baseline. Otherwise, monitoring against irrelevant comparison points.

10. No Segmented Analysis

Monitoring only aggregate performance, missing degradation in specific segments (geography, demographics, product line). Critical issues affecting minorities masked by overall statistics. Always analyze performance by relevant segments.

11. Over-Reliance on Averages

Tracking only average metrics (mean latency, average accuracy) without percentiles or distributions. Averages hide outliers and skew. Use percentiles (P50, P95, P99) and distributions for better understanding.

12. Insufficient Historical Context

Dashboards showing only current state without trends. Cannot determine if current state is normal variance or concerning trend. Always show time series and historical comparison.

Back to Document Index

10. Appendices

Appendix A: Model Monitoring Checklist

Comprehensive checklist for implementing model monitoring:

MONITORING SETUP (Before Deployment)

☐ Model performance metrics defined based on model type
☐ Business metrics identified and aligned with business case
☐ Fairness metrics defined (if applicable)
☐ Baseline performance established from validation/test data
☐ Alert thresholds defined for all critical metrics
☐ Monitoring infrastructure deployed (logging, metrics collection, dashboards)
☐ Automated metric calculation implemented
☐ Dashboards created (operational, performance, executive)
☐ Alerting configured with appropriate severity levels
☐ Escalation procedures documented
☐ Ground truth collection process established
☐ Data drift detection implemented
☐ Model drift detection implemented
☐ Operational health monitoring configured
☐ Monitoring documentation complete

ONGOING MONITORING OPERATIONS

☐ Automated monitoring running 24/7
☐ Dashboards reviewed daily (critical/high models) or weekly (medium/low models)
☐ Alerts triaged and investigated according to severity
☐ Performance metrics calculated on regular cadence
☐ Data drift metrics calculated and reviewed
☐ Fairness metrics monitored
☐ Business metrics tracked and reported
☐ Ground truth data collected and validation performed
☐ Monitoring logs maintained
☐ Incident investigations documented

PERIODIC REVIEWS

☐ Weekly performance review (critical/high models)
☐ Monthly performance review (all models)
☐ Quarterly comprehensive review (all models)

Back to Document Index

☐ Review recommendations documented and tracked
☐ Retraining decisions made based on performance
☐ Baseline updates after retraining
☐ Alert threshold adjustments based on experience
☐ Monitoring effectiveness assessed

MONITORING MAINTENANCE

☐ Alert tuning based on alert history
☐ Dashboard updates to reflect new insights
☐ Monitoring infrastructure scaled with model usage
☐ Deprecated metrics removed
☐ New metrics added as needed
☐ Documentation kept current
☐ Runbooks updated based on incident learnings
☐ Monitoring tested periodically

GOVERNANCE AND COMPLIANCE

☐ Monitoring included in model governance reviews
☐ Monitoring evidence provided for audits
☐ Regulatory monitoring requirements met
☐ Fairness monitoring for regulated uses
☐ Privacy monitoring for personal data
☐ Monitoring reports to governance committees
☐ Audit trail of monitoring and responses maintained

Back to Document Index

Appendix B: Alert Configuration Template

Template for configuring model monitoring alerts:

ALERT CONFIGURATION

Alert Name: _____________
Model: _____________
Alert ID: _____________

Metric Being Monitored:
Metric: _____________
Description: _____________
Calculation Method: _____________

Threshold Configuration:

Critical Threshold:
Condition: [Metric] [Operator] [Value]
Example: Accuracy < 0.70
Severity: Critical
Justification: _____________

High Threshold:
Condition: [Metric] [Operator] [Value]
Example: Accuracy < 0.75
Severity: High
Justification: _____________

Medium Threshold:
Condition: [Metric] [Operator] [Value]
Example: Accuracy < 0.80
Severity: Medium
Justification: _____________

Alert Behavior:
Evaluation Frequency: ☐ Real-time ☐ Hourly ☐ Daily ☐ Weekly
Evaluation Window: _____ (e.g., last 24 hours, last 1000 predictions)
Re-alert Interval: _____ (how often to re-alert if condition persists)
Auto-resolve: ☐ Yes ☐ No
Auto-resolve Condition: _____________

Back to Document Index

Notification Configuration:

Critical Severity:
Channels: □ Page/SMS □ Email □ Slack/Teams □ Incident System
Recipients: _____________
Escalation if not acknowledged in: ____ minutes

High Severity:
Channels: □ Email □ Slack/Teams □ Dashboard
Recipients: _____________

Medium Severity:
Channels: □ Email □ Dashboard
Recipients: _____________

Alert Content:
Title Template: _____________
Message Template: _____________
Include in message:
□ Current metric value
□ Threshold value
□ Trend (last 24 hours, 7 days)
□ Link to dashboard
□ Link to runbook
□ Suggested actions

Response Procedures:
Runbook location: _____________
Investigation steps: _____________
Escalation criteria: _____________
Expected response time: _____________

Alert Metadata:
Created by: _____________
Created date: _____________
Last updated: _____________
Review frequency: □ Quarterly □ Semi-annually □ Annually
Next review date: _____________

Alert History:
Times triggered (last 30 days): ____
False positive rate: ____%
Average time to resolution: ____ hours

Back to Document Index

Tuning notes: ______________

Back to Document Index

Appendix C: Monthly Performance Report Template

Template for monthly model performance reports:

See: Monthly_Performance_Report.xlxs

MONTHLY MODEL PERFORMANCE REPORT

Model Information:
Model Name: ____________
Model Version: ____________
Risk Tier: ☐ Critical ☐ High ☐ Medium ☐ Low
Report Period: ____________
Report Date: ____________
Prepared By: ____________

EXECUTIVE SUMMARY

Overall Health Status: ☐ Healthy ☐ Degraded ☐ At Risk ☐ Critical
Key Findings: ____________
Actions Required: ____________
Recommendations: ____________

PERFORMANCE METRICS

Metric	Current	Baseline	Threshold	Status	Trend
[Metric 1]				✅/⚠️/❌	↑→↓
[Metric 2]				✅/⚠️/❌	↑→↓
[Metric 3]				✅/⚠️/❌	↑→↓

Performance by Segment (if applicable):
[Table showing performance across segments]

Fairness Metrics (if applicable):
[Table showing metrics by demographic group]

DATA DRIFT ANALYSIS

Data Drift Summary:
- Features with significant drift (PSI ≥ 0.25): ____
- Features with moderate drift (PSI 0.10-0.25): ____
- Features with no drift (PSI < 0.10): ____

Back to Document Index

Top Drifted Features:

Feature	PSI Score	Distribution Shift	Impact

MODEL DRIFT ANALYSIS

Prediction Distribution:
- Mean prediction (current): _____
- Mean prediction (baseline): _____
- Shift: _____

Confidence Distribution:
- Predictions with high confidence (>80%): _____%
- Predictions with low confidence (<50%): _____%

OPERATIONAL HEALTH

Availability: _____%
Average Latency: _____ ms (P95: _____ ms)
Error Rate: _____%
Request Volume: _____ (vs. previous month: ↑/→/↓ ____%)
Incidents: _____

BUSINESS VALUE

Business Metrics:

Metric	Current	Target	Status

ROI: _____ (vs. projected: _____)
Stakeholder Feedback: _______________

ALERTS AND ISSUES

Alerts Triggered: _____ (Critical: ___ | High: ___ | Medium: ___ | Low: ___)
Top Alerts:
1. _______________
2. _______________
3. _______________

Open Issues:

Back to Document Index

[List of issues with status and owner]

Resolved Issues:
[List of resolved issues from this period]

ACTIONS AND RECOMMENDATIONS

Immediate Actions Required:
1. ______________
2. ______________

Retraining Recommendation: □ Yes □ No □ Monitor
Justification: ______________

Model Refresh Recommendation: □ Yes □ No
Justification: ______________

Other Improvements: ______________

NEXT STEPS

Scheduled Activities:
- Next performance review: ______________
- Next validation: ______________
- Planned retraining: ______________

APPROVALS

Model Owner: ______________ Date: _____
Reviewed By: ______________ Date: _____

Back to Document Index

Appendix D: Professional Standards Alignment

These Model Monitoring & Performance Standards align with professional standards and frameworks:

BABOK (Business Analysis Body of Knowledge) Alignment:

• Solution Evaluation (6): Evaluating solution performance and value delivery
• Requirements Analysis and Design Definition (5): Defining performance requirements and acceptance criteria
• Requirements Life Cycle Management (3.1): Maintaining and updating requirements based on performance
• Business Analysis Planning and Monitoring (3): Planning for performance monitoring and reporting

PMBOK (Project Management Body of Knowledge) Alignment:

• Project Quality Management (8): Monitoring quality and performance
• Project Risk Management (11): Monitoring risks and risk responses
• Project Communications Management (10): Performance reporting to stakeholders
• Project Integration Management (4): Integrating monitoring across all project areas
• Monitor and Control Project Work: Tracking, reviewing, and reporting progress

DMBOK (Data Management Body of Knowledge) Alignment:

• Data Quality Management (13): Monitoring data quality continuously
• Data Operations (14): Operational data management and monitoring
• Metadata Management (12): Documenting monitoring metadata and lineage
• Data Governance (3): Governance of data and model performance
• Data Security (7): Monitoring security and privacy compliance

Model Risk Management Standards:

• SR 11-7 (Supervisory Guidance on Model Risk Management): Ongoing monitoring requirement for model risk management
• Federal Reserve SR 13-19: Guidance on Managing Outsourcing Risk (vendor model monitoring)
• OCC 2011-12: Supervisory Guidance on Model Risk Management
• Model Validation Standards: Ongoing validation and performance monitoring
• Basel Committee on Banking Supervision: Model validation and monitoring principles

Back to Document Index

AI Governance Standards:

- NIST AI Risk Management Framework: Continuous monitoring and measurement
- ISO/IEC 42001:2023 (AI Management System): Monitoring and measurement requirements
- ISO/IEC 23894 (AI Risk Management): Continuous risk monitoring
- ISO/IEC 5338 (AI System Life Cycle): Operational monitoring requirements
- EU AI Act: Monitoring obligations for high-risk AI systems
- OECD AI Principles: Robustness and accountability through monitoring

DevOps and SRE Practices:

- Site Reliability Engineering (SRE): Service level objectives, monitoring, and alerting
- DevOps Monitoring Best Practices: Continuous monitoring in production
- MLOps: Machine learning operations and monitoring practices
- Observability: Metrics, logs, traces for system understanding
- Infrastructure as Code: Monitoring configuration as code

Quality and Process Standards:

- ISO 9001 (Quality Management): Monitoring and measurement of processes and outputs
- CMMI (Capability Maturity Model Integration): Process monitoring and control
- Six Sigma: Statistical process control and monitoring
- Lean: Continuous monitoring and improvement
- ITIL (IT Service Management): Service monitoring and performance management

Back to Document Index

Document Usage Rights and Disclaimer

This Business Analysis Template for AI Projects is provided as a starter document to assist business analysts in assessing organizational readiness for AI adoption.

Usage Rights:

✓ You may freely use, modify, and customize this template for your projects

✓ You may adapt the readiness assessment framework to fit your needs

✓ You may incorporate this into your organizational assessment processes

✓ You may share this template within your organization

Restrictions:

✗ You may not resell, redistribute, or commercialize this template

✗ You may not claim original authorship of this framework

✗ You may not remove these usage rights statements

Disclaimer:

This template is provided as-is without warranties. While it incorporates professional best practices for readiness assessment and organizational change management, users are responsible for adapting methods to their specific context. The template should be customized based on your unique needs and validated with appropriate subject matter experts including change management professionals, organizational development specialists, and business leaders. Readiness assessment requires understanding of both technical capabilities and organizational dynamics.