**Alipay Fraud Model Monitoring Plan**

*(For Model Risk Management Submission)*

**1. Executive Summary**

The Alipay fraud model is a real-time, machine learning-driven system that assigns a fraud risk score to each transaction. The purpose of this monitoring framework is to assess the model's predictive performance, stability, operational effectiveness, and compliance with regulatory expectations under SR 11-7 and internal model risk governance standards.

**2. Scope of Monitoring**

* Transaction-level fraud detection scoring
* Decision outputs: Approve, Decline, Manual Review
* Real-time fraud detection performance
* Stability of model behavior over time

**3. Core Outputs Monitored**

* **Risk Score** (continuous value per transaction)
* **Predicted Decision** (Approve, Decline, Manual Review)
* **Disposition/Actual Outcome** (Confirmed Fraud / Legitimate)
* **Reason Codes / Score Drivers** (Top features impacting decision)

**4. Data and Variables Required**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Usage in Monitoring** | **Business Significance** |
| **Transaction ID** | Unique identifier to trace and audit each transaction. | Enables record-level tracking and reconciliation. |
| **Timestamp** | Capture the transaction date and time for time-based analyses. | Helps in trend monitoring, seasonality detection, and latency measurement. |
| **Risk Score** | Used for ROC, AUC, PSI calculation, score distribution analysis. | Primary output that represents predicted probability of fraud. |
| **Predicted Decision** | Model's action (Approve / Decline / Manual Review) | Helps understand model operational behavior; used for confusion matrix creation. |
| **Actual Disposition (Fraud/Legitimate)** | Real outcome based on manual investigation or post-transaction behavior. | Enables performance calculation (Precision, Recall, Accuracy). |
| **Reason Codes** *(optional but recommended)* | Main drivers behind risk score assignment. | Important for business insights, model transparency, fraud typology shift detection. |
| **Latency (Milliseconds)** | Time between transaction event and model decision. | Critical for real-time model SLA monitoring and user experience. |

**5. Key Monitoring Metrics and Formulas**

**5.1 Confusion Matrix Metrics**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Formula** | **Interpretation** |
| **True Positives (TP)** | Predicted Fraud and Actually Fraud | Model correctly identifies fraud |
| **True Negatives (TN)** | Predicted Legitimate and Actually Legitimate | Model correctly approves good transactions |
| **False Positives (FP)** | Predicted Fraud but Actually Legitimate | Model incorrectly flags legitimate users |
| **False Negatives (FN)** | Predicted Legitimate but Actually Fraud | Model misses fraudulent transactions |

**Derived Metrics:**

* **Precision** = TP / (TP + FP)  
  *Proportion of flagged frauds that were actual frauds.*
* **Recall (Capture Rate)** = TP / (TP + FN)  
  *Ability to capture total fraud cases.*
* **Accuracy** = (TP + TN) / (TP + TN + FP + FN)  
  *Overall correctness of the model.*
* **False Positive Rate (FPR)** = FP / (FP + TN)  
  *Proportion of legitimate transactions wrongly flagged.*

**5.2 AUC-ROC Curve**

* **Plot:** Sensitivity (Recall) vs 1-Specificity (FPR) at various thresholds.
* **Metric:** Area under the ROC curve (AUC).
* **Interpretation:**
  + AUC = 0.5 → Random model
  + AUC = 1.0 → Perfect model
  + AUC > 0.80 → Acceptable discrimination power.

**5.3 Population Stability Index (PSI)**

* **Formula:**

PSI=∑((Actual%−Expected%)×ln⁡(Actual%Expected%))\text{PSI} = \sum \left( (Actual\% - Expected\%) \times \ln\left(\frac{Actual\%}{Expected\%}\right) \right)PSI=∑((Actual%−Expected%)×ln(Expected%Actual% ))

* **Interpretation:**
  + PSI < 0.10 → Stable
  + 0.10 < PSI < 0.25 → Moderate drift (monitor)
  + PSI > 0.25 → Significant drift (investigate)

**5.4 Threshold Monitoring**

* Distribution of transactions across Approve, Decline, and Manual Review bands.
* Metric: % Approve / % Decline / % Manual Review
* **Business Significance:**
  + Helps detect threshold misalignment if fraud trends shift.
  + Ensures operational loads on manual review teams remain manageable.

**5.5 Operational Metric: Latency**

* **Latency = (Model Decision Time - Transaction Initiation Time) in milliseconds.**
* **Target:**
  + Fraud decision should occur within 500 milliseconds to meet real-time requirements.
* **Business Impact:**
  + High latency impacts user experience and transaction completion rate.

**6. Monitoring Frequency**

* Confusion Matrix Metrics: **Monthly**
* AUC and ROC Curve: **Monthly**
* PSI (Risk Score Stability): **Quarterly**
* Threshold and Operational Metrics: **Monthly**

**7. Governance and Escalation**

* **AUC deterioration below 0.75** → Escalation to MRM and Model Owner.
* **False Positive Rate increase >5%** → Investigation and impact assessment.
* **PSI >0.25** → Recalibration or redevelopment discussion.
* **Latency SLA breach** → Escalation to Technology Risk and Vendor Management.

**8. Data Requirements Summary Table**

|  |  |  |
| --- | --- | --- |
| **Data Element** | **Required Format** | **Frequency** |
| Transaction ID | Alphanumeric | Per Transaction |
| Timestamp | YYYY-MM-DD HH:MM:SS | Per Transaction |
| Risk Score | Numeric (0-1000 or 0-1) | Per Transaction |
| Predicted Decision | Categorical (Approve, Decline, Manual Review) | Per Transaction |
| Disposition | Categorical (Fraud, Legitimate) | Per Transaction |
| Reason Codes | Text / Multiple Labels | Per Transaction |
| Latency | Milliseconds (Integer) | Per Transaction |