Model Monitoring Pipeline

A model monitoring pipeline is essential for ensuring sustained model performance and reliability in production. It should encompass several components: data collection, storage, performance monitoring, drift detection, alerting, and visualization.

Data Collection and Storage

Three types of data should be collected and stored:

1. Model Predictions: Real-time or batch predictions generated during model serving.
2. Input Data Features: The input data fed into the model during inference.
3. Ground Truth Labels (if available): Actual outcomes collected post-inference, which serve as the benchmark for model performance evaluation.

These datasets should be stored in a database, with appropriate identifiers to enable merging of these data for analysis.

Performance Monitoring

Performance metrics appropriate to the model type should be computed regularly (e.g. daily, weekly), with the cadence depending on the use case. For classification models, metrics like accuracy, precision, recall, F1-score, and AUC-ROC are relevant. For regression models, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared are commonly used. These metrics should be automatically calculated on both overall and segmented data to identify performance issues and potential bias.

Drift Detection

There are different types of model drift and ways to detect them. Drift detection should be scheduled at regular intervals.

1. Data drift (or covariate drift) occurs when the distribution of input features is different from what was seen during model training. This shift in input data distribution can lead to worsened model performance as the model cannot generalize beyond what it has seen during training. Statistical tests such as the Kolmogorov-Smirnov test can be used to compare the distribution of new data against the distribution of training data. Its p-value can be used to determine if drift has occurred.
2. Concept drift occurs when there are changes in the relationship between input features and the target variable. Monitoring performance metrics can help detect this drift. To handle concept drift, the model would need to be retrained using new data.

3. Prediction drift occurs when there are deviations in the distribution of model outputs. It is an indication that a change has happened in the input data which affected the model's predictions.

Alerting Mechanisms

An automated alerting system should be implemented to notify stakeholders (e.g. data scientists and engineers) when model performance degrades or drift is detected. Thresholds for alerts can be based on historical performance metrics. When performance degrades beyond the threshold or significant drift occurred beyond the threshold, notifications can be automatically sent via emails or messaging platforms.

A retraining pipeline can be also triggered using updated data. This retraining can be automated or manually reviewed, depending on business requirements.

Visualization

A real-time dashboard should visualize key performance indicators (KPIs) and drift metrics. This dashboard could include:

- Time series plots for model performance metrics over time.
- Histograms and box plots showing the distribution of input features and predictions (compared against what was observed during model training).
- A summary of alerts and notifications.