

THE DEFINITIVE Machine Learning Observability Checklist for Fintech

What to look for when assessing ML monitoring and observability platforms

Introduction

The fight for the future of the financial services industry is heating up. In the wake of COVID-19, fintech companies are changing how financial services are structured, delivered and consumed with remarkable speed. AI is playing a starring role, influencing business-critical decisions on everything from customer experience to fraud prevention and lending.

Despite this progress, 28.6% of ML teams in the financial services industry say it still takes them a week more to detect and fix an issue with a model in production – and 50.6% report that their jobs are harder than before the pandemic due to elevated drift and performance degradation. How can fintech companies and others in the industry stay a step ahead?

ML observability is a key part of the answer and a **critical element for any modern ML infrastructure stack**. This checklist covers the essential elements to consider when evaluating an ML observability platform. Based on our team's firsthand experience building ML teams from the ground up and tracking billions of daily predictions on behalf of financial services companies, this buyer's blueprint is designed to inform product and technical requirements for RFPs or individual vetting.

Model Lineage, Validation & Comparison

- Model versioning and lineage support
- Pre-Launch model validation

Data Quality Monitoring & Troubleshooting

- Monitor production model for bad inputs
- Configurable real-time statistics on features & predictions (min, max, median, mean, standard deviation) in aggregate and by cohorts
- Ability to detect anomalous behavior (outlier detection) on predictions
- Configurable baseline setup

Drift Monitoring & Troubleshooting

- Overall production drift detection (concept, data, model)
- Compare training versus production distributions
- Drift monitoring on any flexible dataset
- Drift detection across any cohort
- Troubleshooting model drift by drilling into feature drift
- Configurable baseline setup

Performance Monitoring & Troubleshooting

- Monitor ground truth by combining predictions with delayed response label data
- Production A|B comparison of models
- Configurable baselines that support both production and pre-production
- Ability to compare model performance metrics (such as ROC-AUC, PR-AUC, accuracy, precision, recall, r-squared, MSE, MAE) from trained model to production model (or two other periods of time)
- Monitor production models using constant thresholds and dynamic thresholds
- Automatically surface up performance problems by feature, value or cohort without a user needing to write SQL queries
- Ability to perform dynamic cohort analysis/segmentation of predictions
- Dashboards that non-technical stakeholders can understand

Explainability

- Ability to view the feature importance for the top n features
- Support for global, cohort, and local explainability to assist in all stages of ML lifecycle
- Does not require a model upload to use explainability

Business Impact Analysis

- Custom user defined function (UDF) to tie model performance back to business metrics
- Dynamically analyze thresholds for probability-based decision models
- Compare pre-production models to current production models — champion and challenger

Integration Functionality

- Agnostic of model types/libraries
- Support SaaS, on-prem and hybrid deployments
- Specializes in model monitoring and observability instead of providing an end-to-end hosting and serving system (ensures product depth, user customization and choice)
- Ability to set up alerts that integrate with PagerDuty or your preferred incident response platform
- Automatically infers the model type and calculates the appropriate metrics for monitoring
- Ability to easily import data from and export to external data sources

UI/UX Experience

- Flexible, customizable dashboards that technical and non-technical stakeholders can check to determine if models have changed
- Dark mode (optional)

Scalable To Meet Current & Future Analytics Complexity

- Ability to handle analytic workloads (analytic data housed in an OLAP)
- Ability to support load testing (>500 features, >50M predictions per day to total >1BN over testing period)

Built To Meet the Needs of Fintech & Financial Services Companies

- Automatically surfaces problematic segments (i.e. high false negative rate), visualizing blindspots via heatmaps
- Ability to use proxy metrics like drift to navigate delayed or significantly delayed actuals
- Ability to submit protected classes (i.e. age, gender) as features to a model without using them in the model itself for bias/fairness checks, governance and compliance
- Configurable pre-production or production baselines, the latter being especially important in fraud models
- Prediction drift impact measurement to surface what features have had significant drift impact

Request a Trial

Newsletter Signup