WHITEPAPER

Ensuring Optimal Performance and Integrity of ML Models

Why and how to monitor machine learning models

fractal

Introduction

Businesses are increasingly dependent on data-driven decision-making to stay ahead of the competition. With the power to harness massive volumes of data, machine learning (ML) models allow businesses to make intricate predictions, enhancing their capacity to provide exceptional products and services catering to the ever-evolving needs of customers and the market.

However, after their training and deployment, these ML models are not isolated entities: they operate within real-world data, which constantly fluctuates and progresses, significantly influencing the models' effectiveness. Consequently, the accuracy of predictions declines, potentially leading to suboptimal business decisions and an incomplete understanding of market dynamics.

Unlike traditional software, where post-deployment monitoring usually focuses on uptime and error rates, ML models require monitoring strategies that account for their unique nature. This monitoring needs to go beyond technical performance, focusing instead on how the model interacts with real-world data and how effectively it makes accurate predictions over time. This is where the practice of 'model monitoring' (MM) becomes crucial.

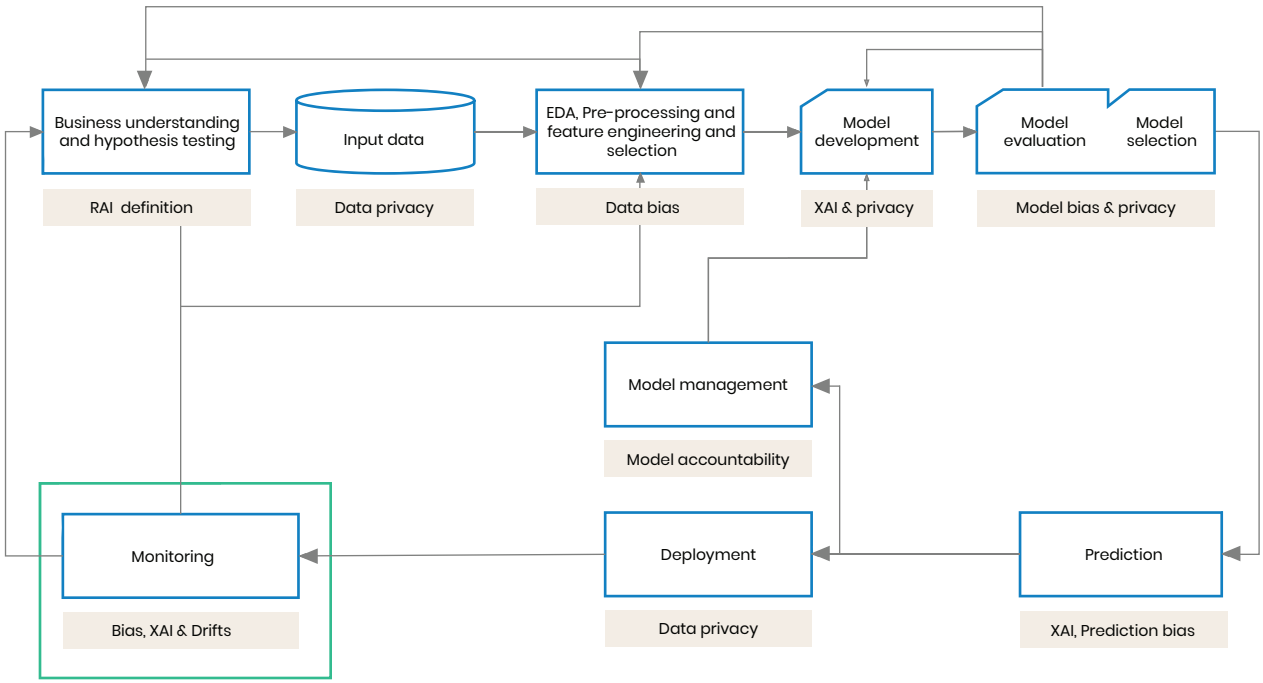
Model Monitoring

Model monitoring is the process of consistently overseeing the performance of ML models post-deployment. It involves tracking key metrics, assessing input and output data, detecting shifts in data trends, and ensuring that the model continues to make accurate predictions in line with the business objectives.

Accelerate responsible business performance with model monitoring

MM is a crucial aspect of the ML lifecycle. It involves an ongoing process of evaluating the performance of deployed ML models, going beyond mere prediction accuracy. It encompasses a comprehensive system of checks and balances to safeguard the model's continued performance, validity, and reliability over time. Model monitoring is pivotal in maintaining ML models' peak performance and maximum benefit delivery for businesses.

Where and how model monitoring factors into responsible AI performance



Drift and model accuracy

Drift refers to the phenomenon where the performance of an ML model degrades over time due to changes in the input data distribution. It occurs when the patterns or characteristics of the data used for training the model differ from those observed in the data the model encounters in the real world. The data distribution shift can compromise the model's accuracy and effectiveness, hindering its ability to make precise predictions on unfamiliar data.

There are four main categories of drift, each with its unique causes and challenges.

Concept drift

This occurs when the statistical characteristics of the variable being predicted by the model undergo changes over time. Real-world scenarios, market trends, and customer behaviors are dynamic and can vary. When the patterns a model learns during training no longer apply, the model's accuracy can decrease.

Anomalies and outliers

Identifying anomalies or outliers in data or model predictions can indicate problems such as pipeline bugs, data distribution shifts, or model issues. Detecting these anomalies is challenging, especially in complex data or when the model's standard behavior is unclear.

Data drift

Data drift is similar to concept drift, but rather than the target variable, it refers to the shifts in the input data's statistical properties over time. If the model's input data changes significantly from what it was trained on, it can lead to inaccurate predictions.

Model decay

Even if the underlying data remains consistent, a model's performance can degrade over time. This could be attributed to alterations in the surrounding environment, affecting the model's operations, or to untracked variables.

Tracing the root cause of drift

To find the source of drift, MM needs to consider several factors, namely feature importance, data input, and bias.

FEATURE IMPORTANCE

Feature importance in ML models can change for various reasons, such as evolving data patterns or alterations in relationships between variables. It can also be influenced by changes in data sources, like the unavailability or modification of previously essential variables. Ensuring the accurate detection of concept drift is vital for optimal monitoring of feature importance. Concept drift refers to the unexpected fluctuation in the statistical properties of the target variable, that can significantly impact the model's predictive performance.
Regular monitoring of feature importance helps identify these changes early, enabling adjustments, retraining, or model rebuilding before significant performance degradation occurs. Techniques like permutation importance, SHAP values, or LIME can be employed to understand and monitor feature importance effectively.

DATA INPUT

Data input monitoring is a vital aspect of MM, focusing on detecting changes in input data that can disrupt original training assumptions that can otherwise lead to model drift, bias, accuracy decline, and privacy concerns.The core of data input monitoring lies in regularly evaluating real-world data to identify shifts in feature distribution. It goes beyond noting changes and aims to understand their potential impact, signaling the need for further analysis or model adjustments. Measuring the probability of changes in data distribution and their effect on outcomes is crucial for identifying accuracy, fairness, and reliance on specific features. This process is similar to analyzing model performance without having access to ground truth data.While accounting for potential dependencies among features can be complex, data input monitoring often treats each feature as independent. This simplifies the analysis while providing an understanding of drift magnitude.

BIAS AND FAIRNESS

ML models are influenced by the data they learn from. If the training data contains biases, the models may replicate and perpetuate these biases. Ensuring fairness in model predictions is essential to prevent unfair treatment of certain groups, which can have unethical or illegal consequences. Fairness monitoring is crucial to observe model predictions and prevent disproportionate disadvantages for specific groups or individuals. Bias detection identifies and understands inherent biases within the model or its training data. It includes analyzing input features, model decisions, and outcomes to prevent any systematic bias or discrimination. Recognizing that bias may appear subtly, such as through seemingly innocuous features that substitute for prohibited attributes is crucial.

When drift is detected

Ensuring AI models' ongoing performance and reliability in real-world scenarios requires diligent monitoring and addressing of drift. Unfortunately, there is no silver bullet solution.

There are several methods for model accuracy monitoring, with some common ones being:

Retraining with new data	The model is retrained periodically using recent data, and its predictions are compared to the actual outcomes to gauge accuracy. This process often includes re-tuning hyperparameters.
Online evaluation	In this real-time evaluation, the model's predictions are compared to actual outcomes as they occur. This method requires a mechanism for capturing predictions and outcomes in real time.
Offline evaluation	This method involves maintaining a holdout data set to evaluate the model periodically. This data set should reflect recent changes in the data distribution.

Model monitoring framework

An optimal MM framework should include components such as:

Comprehensive tracking

An effective MM solution should track all aspects of a model's performance, including predictive accuracy, distribution of predictions, data drift, and concepts. The monitoring system should also track metadata, such as model versions, hyperparameters, and training data, to comprehensively understand model behavior over time.

Integration and interoperability

An MM solution should seamlessly integrate with existing ML platforms and infrastructure. It should also support interoperability across programming languages, ML frameworks, and data storage systems.

Real-time alerts and reporting

An MM solution should have configurable alerts based on performance metrics or significant changes in model behavior. Furthermore, it should facilitate regular reporting, enabling review and analysis of the model's performance over time.

Scalability and efficiency

The MM solution must be able to handle multiple models operating concurrently, effectively manage substantial data volumes, and seamlessly expand to accommodate an increasing number of models or data.

Model explainability

The MM solution should have integrated tools for model explainability, such as SHAP or LIME, to provide insights into how models make decisions.

Bias and fairness monitoring

It should provide mechanisms to check for disparate impact, model fairness, and other relevant ethical considerations across different demographics or groups.

Feedback loop management

This helps prevent model performance from spiraling downwards and allows human intervention using HITL.

Security and compliance

The MM solution should provide encryption, access controls, and audit logs, among other security features.

By incorporating these crucial aspects, an MM solution empowers organizations to proactively monitor, detect, and mitigate potential issues, enabling them to make informed decisions and maintain the integrity of their ML initiatives.

Applying MM in the real world — A use case example

Imagine you have an ML model that predicts the sales volume of a particular product based on several features, one of which is the product's price. This model has been trained on historical data where the product price was relatively stable. In response to external factors such as rising raw material costs, shifts in market competition, and inflation, the product's price experiences notable fluctuations over time. Consequently, the correlation between price and sales volume deviates from the pattern observed in the training data, illustrating the concept of concept drift.

Left unchecked, the model's sales volume predictions may become inaccurate, leading to poor business decisions and potentially significant revenue loss. Therefore, detecting this drift in the data (in this case, the change in product price) is critical. Once the drift is detected, it signals that the predictive model may need to be updated or retrained to maintain its accuracy.

Discovering drift detection methods entails recognizing alterations in data distribution, such as employing the KS test, utilizing ML methods capable of detecting distribution changes, or employing straightforward techniques like monitoring the mean or variance of the price. Subsequently, an alert will be triggered if significant changes are detected. Standard statistical tests employed in this scenario include the KS test, Page-Hinkley, PSI/PPS, and CSI.

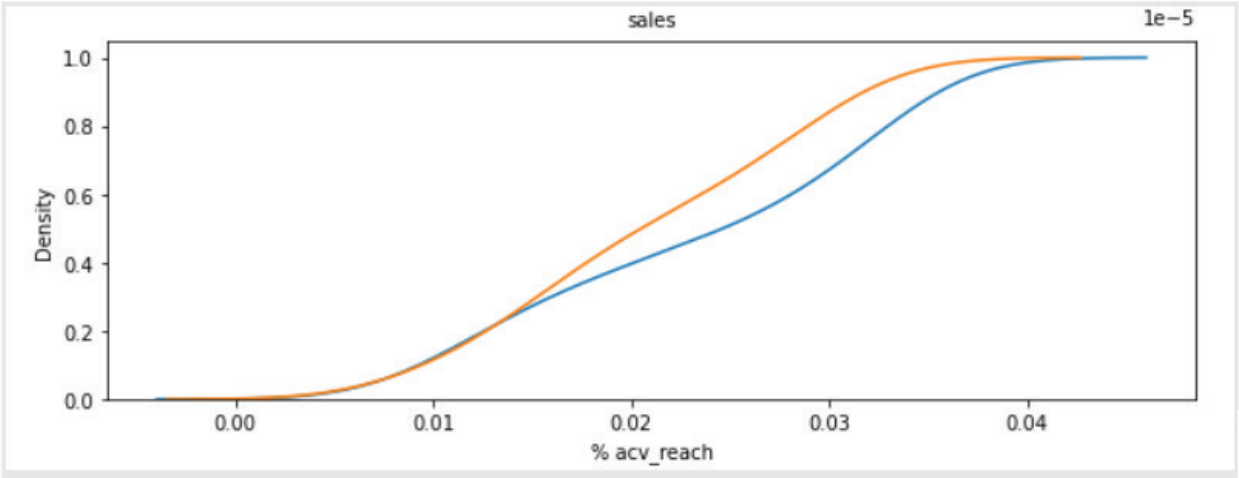


Figure 2. KS Test

Suppose the KS test indicates a significant difference between the two data sets (e.g., Actual vs. Predicted Sales). In that case, we might use this information to identify areas where the enterprise needs to improve its sales strategy, such as by targeting different customer segments or changing its marketing approach.

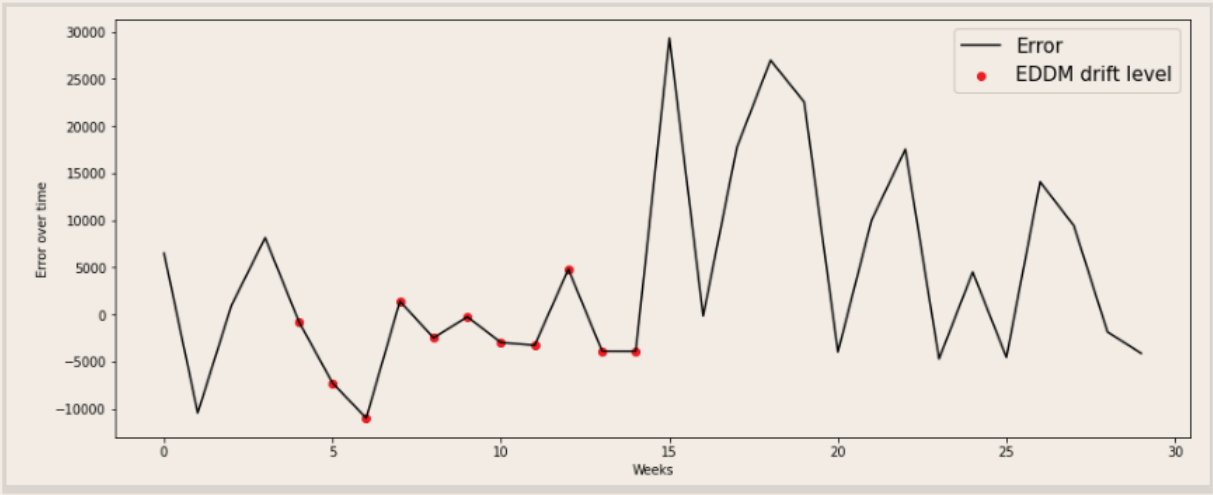


Figure 3. Page-Hinkley Test

The **Page-Hinkley Test** generates a score indicating the drift or change level in the data stream. If this score exceeds a certain threshold, it indicates a change in the data stream.

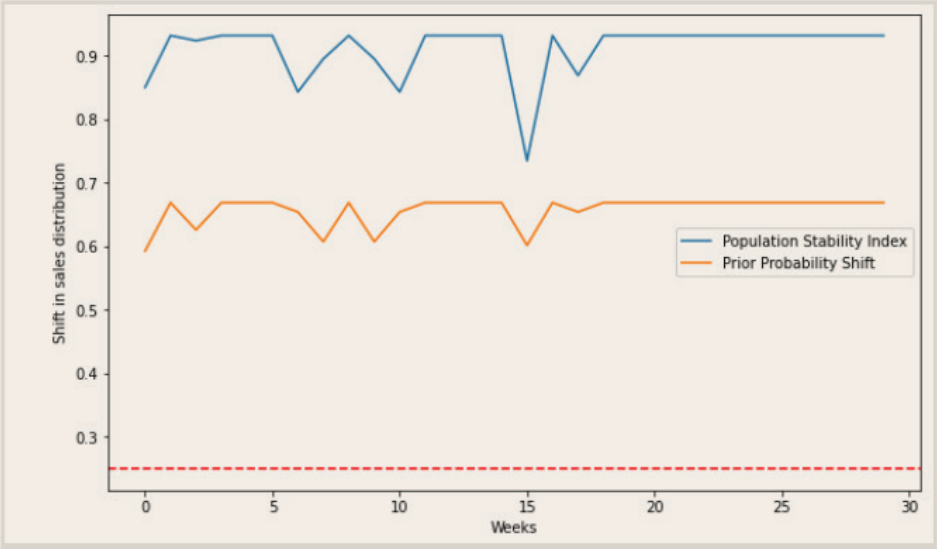


Figure 4. PSI and PPS test

Population Stability Index (PSI) evaluates the consistency between the distribution of a scoring variable (predicted probability) in a scoring data set and the training data set used to build the model. The purpose is to assess the similarity between the current scoring and the predicted probability derived from the training data set. **Prior Probability Shift (PPS)**, on the other hand, occurs when the distribution of the target variable changes while the distribution of the input features remains unchanged.

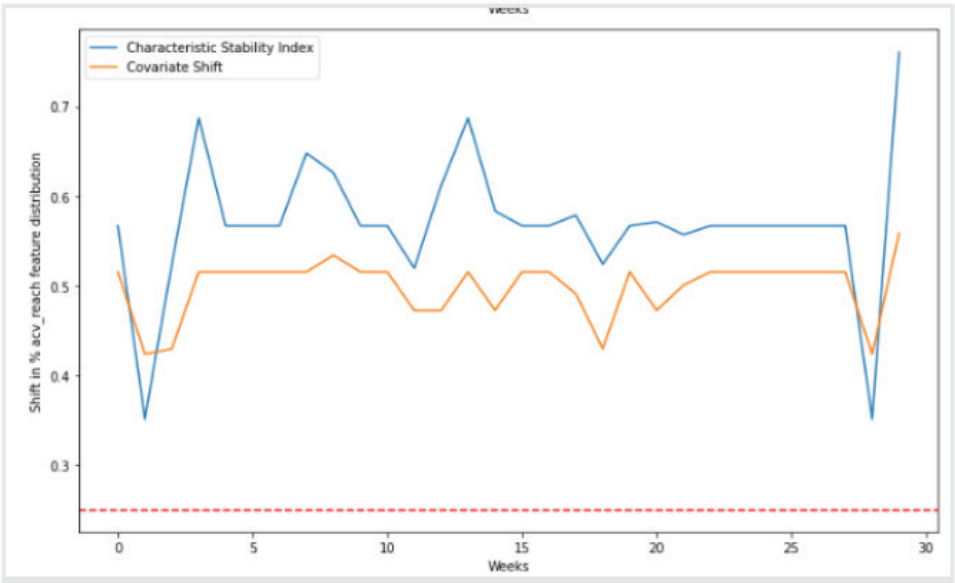


Figure 5. CSI test

Characteristic Stability Index (CSI) is a measure of the stability of the distribution of the corresponding feature. In the figure above, the blue line shows CSI, and the orange line shows the **Covariate Shift (CS)**. The plots show the results of monitoring the stability of the distributions of the independent variables over time.
The red line represents the warning threshold, which is a value of 0.25. If the CSI or CS exceeds the warning threshold, it indicates that there may be a problem with the stability of the corresponding feature, and further investigation may be necessary.

Conclusion

In today's dynamic data-driven world, maintaining the accuracy, reliability and value of machine learning models is crucial. Businesses need a robust monitoring framework to navigate the complexities and challenges of model drift, feature importance, and bias detection.

Businesses can ensure accurate model predictions, address drift, monitor feature importance, and ensure fairness through a robust monitoring framework. By leveraging advanced techniques and continuous monitoring, enterprises can uphold the integrity and performance of their ML models, driving better decision-making and improved outcomes.

Author

Olena Fylymonova
Consultant,
Strategic Center

About Fractal

Fractal is one of the most prominent providers of Artificial Intelligence to Fortune 500® companies. Fractal's vision is to power every human decision in the enterprise, and bring AI, engineering, and design to help the world's most admired companies.

Fractal's businesses include Crux Intelligence (AI driven business intelligence), Eugenie.ai (AI for sustainability), Asper.ai (AI for revenue growth management) and Senseforth.ai (conversational AI for sales and customer service). Fractal incubated Qure.ai, a leading player in healthcare AI for detecting Tuberculosis and Lung cancer.

Fractal currently has 4000+ employees across 16 global locations, including the United States, UK, Ukraine, India, Singapore, and Australia. Fractal has been recognized as 'Great Workplace' and 'India's Best Workplaces for Women' in the top 100 (large) category by The Great Place to Work® Institute; featured as a leader in Customer Analytics Service Providers Wave™ 2021, Computer Vision Consultancies Wave™ 2020 & Specialized Insights Service Providers Wave™ 2020 by Forrester Research Inc., a leader in Analytics & AI Services Specialists Peak Matrix 2022 by Everest Group and recognized as an 'Honorable Vendor' in 2022 Magic Quadrant™ for data & analytics by Gartner Inc.

For more information, visit fractal.ai

Corporate Headquarters
Suite 76J,
One World Trade Center, New York,
NY 10007

Get in touch