# Mitigating AI Model Drift with IBM Watson OpenScale

By Manish Bhide
Chief Architect, Watson OpenScale

Airline pilots have a major responsibility: They must fly a plane full of passengers from one city to another safely. To ensure they are ready to meet this challenge, they must pass periodic reviews of their health, flying skills, and ability to perform emergency procedures.

Similarly, AI models in a growing number of organizations have a major responsibility. They are entrusted to help make key business decisions such as predicting who should receive a loan, a life insurance policy, or a specific marketing campaign.

Before AI models are deployed, they need to demonstrate they that make accurate predictions on test data. After they are deployed, they should be periodically tested—just like airline pilots— to verify they remain able to fulfill their responsibilities. The problem for AI models is that as they encounter new data in production, they change, and their accuracy may deteriorate.

One way to test a production model's accuracy is with new manually labelled test data. However, labeling test data is expensive and time-consuming. Only a limited amount can usually be created, and it may not replicate the variety of data that a model encounters in production. Therefore, manually labelled test data is not adequate to ensure production model accuracy.

## Automatically detecting drift

To help organizations confidently run their AI and ML models, IBM Watson OpenScale has a feature called drift detection. It enables organizations to test AI models in production without requiring manually labelled test data.

Here is how it works: Poor model accuracy often results when an AI model encounters production data that differs from the original training data. For instance, a model that predicts whether home loan applications should be approved may have been trained on home prices in a certain range. While a model is in production, home prices may increase by 25 percent. This can cause the model to start making incorrect predictions.

To protect against this, the drift detection feature in IBM Watson OpenScale monitors the data a production model receives, and it estimates the accuracy of the model's outputs. It generates an alert if accuracy drops below a desired threshold. This can help organizations mitigate reduced model accuracy before it significantly affects business outcomes. For instance, a home loan approval model can be re-trained and corrected before it approves borrowers that are outside the designated risk profile.

## Configuring drift detection in OpenScale

When users configure drift detection in OpenScale, they specify a tolerable drift magnitude. Drift is measured as the decrease in model accuracy compared to accuracy at training time. If accuracy during training was 90 percent and accuracy during runtime is estimated at 80 percent, then the model is said to have drifted by 10 percent.

Depending on the use case, model owners are willing to tolerate different amounts of drift. To support these differences, IBM Watson OpenScale allows each model owner to specify the drift magnitude (called the drift alert threshold) for their model. If drift for a model rises above the specified threshold, OpenScale generates an alert.

## How does drift detection work?

To identify drift, OpenScale needs to understand the behavior of the user's model on training and test data. It analyses model behavior and builds its own model (called the drift detection model) which predicts if the user's model is going to generate an accurate prediction for a given data point. OpenScale runs the drift detection model on the payload data (the data that the model receives during runtime). It identifies which records in the payload are likely to generate a prediction error in the user's model, and it calculates the overall accuracy of the model on that payload. This value is compared with the accuracy of the model at training time, and the difference is

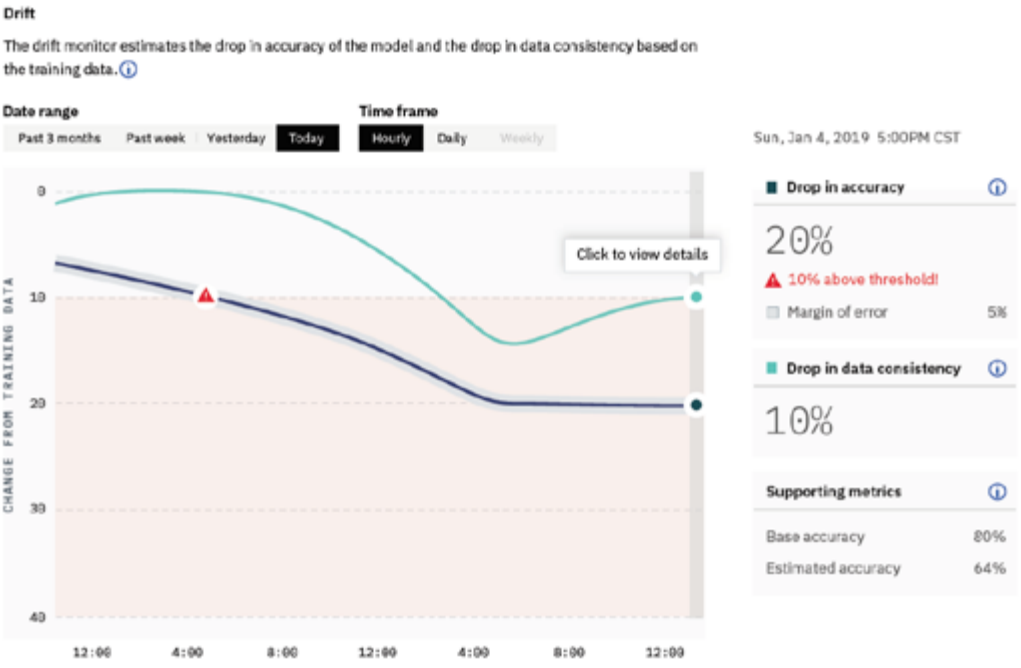
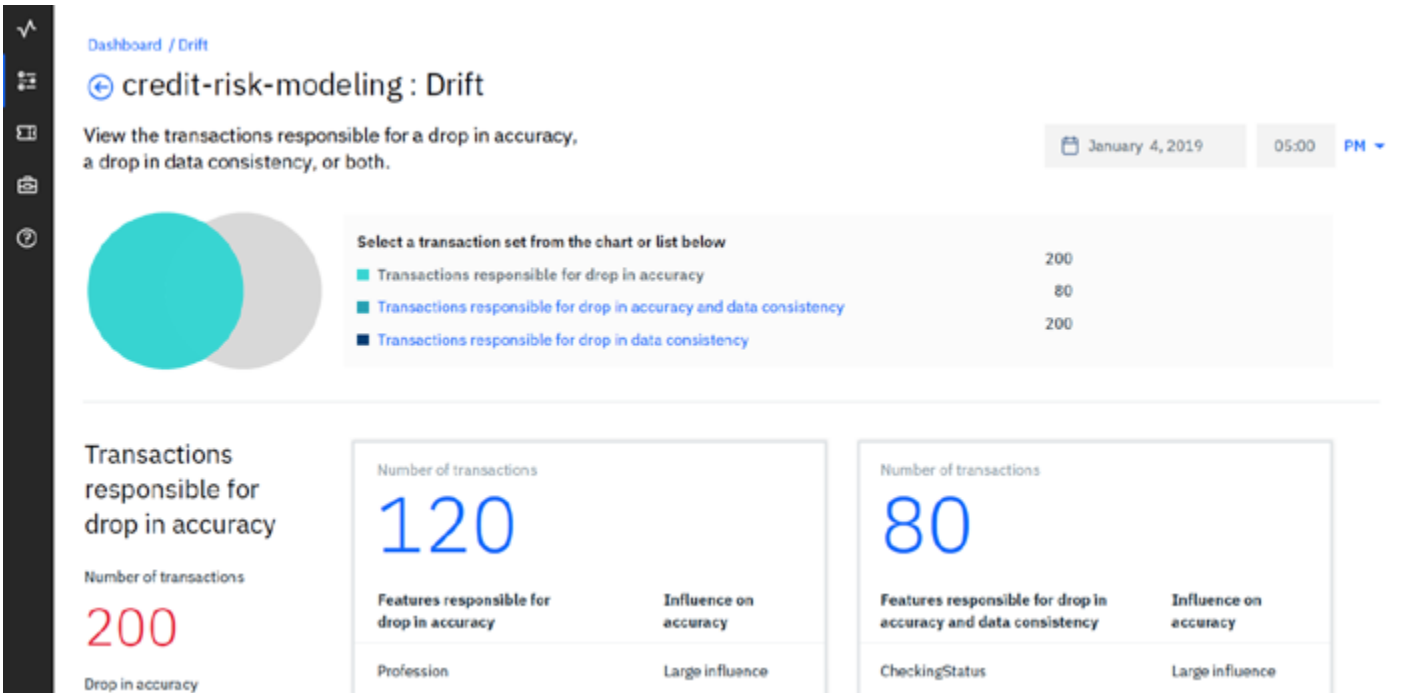



*Figure 1*: Drift detected by IBM Watson OpenScale has risen to 20 percent.

## Reporting drift

Watson OpenScale's drift detection interface, shown in Figure 1, tracks drift over time. The drift indicated in Figure 1 is 20 percent, ten percent above the threshold set by the user. Drift is represented as a thick grey band in the graph, with a dark center line. The center line represents predicted accuracy. However, finding the exact predicted accuracy is a challenging problem, and therefore, the grey band reflects a range within which the actual drift is likely to lie. The model is said to have drifted if the predicted accuracy (center of the band) crosses above the pre-set threshold.




model drift.

*Figure 2*: Watson OpenScale groups transactions causing drift by the top features responsible for it.

## How to fix AI model drift

Once drift has been identified, the user can drill down to a point in time to understand the transactions leading to drift. Figure 2 shows that OpenScale has classified the transactions causing drift into three groups, characterized by the most prominent feature values in each group that led to drift. The left-most box in Figure 2 represents 120 drift-causing transactions, typified by certain values for profession and state. OpenScale has identified these values as key reasons why the model is likely to make an incorrect prediction for these transactions. Please note that the particular values shown in each box represent only one of multiple transactions in the group. Clicking on the box reveals a complete list of transactions in that group.

As a first step in mitigating drift, a user can send these transactions for manual labelling. This manually labelled data can be used to retrain the model so that model accuracy does not drop at runtime. In this way, IBM Watson OpenScale is designed to not only help identify drift, but also highlight its root cause and provide transactions which can be turned into training data and used to fix the model's drift.

## Keep your AI models accurate

To sum up, the accuracy of an AI model's predictions can change when production data differs from training data. Organizations need to monitor their models as they encounter production data to ensure that accuracy stays at expected levels. The drift detection feature of IBM Watson OpenScale is designed to help organizations track model accuracy at runtime, and it identifies production data that differs enough from training data that it is causing excess drift. These transactions can used to re-train the model and maintain its accuracy. Watson OpenScale helps ensure that enterprises can quickly react to a change in model accuracy before it has significant impact on business outcomes. OpenScale drift detection, along with its bias detection and explainability features, provide the guardrails that organizations need to keep AI updated and take models from pilot to production.

Get the big picture, and get started with Watson OpenScale for free.



IBM Global Services
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
August 2019