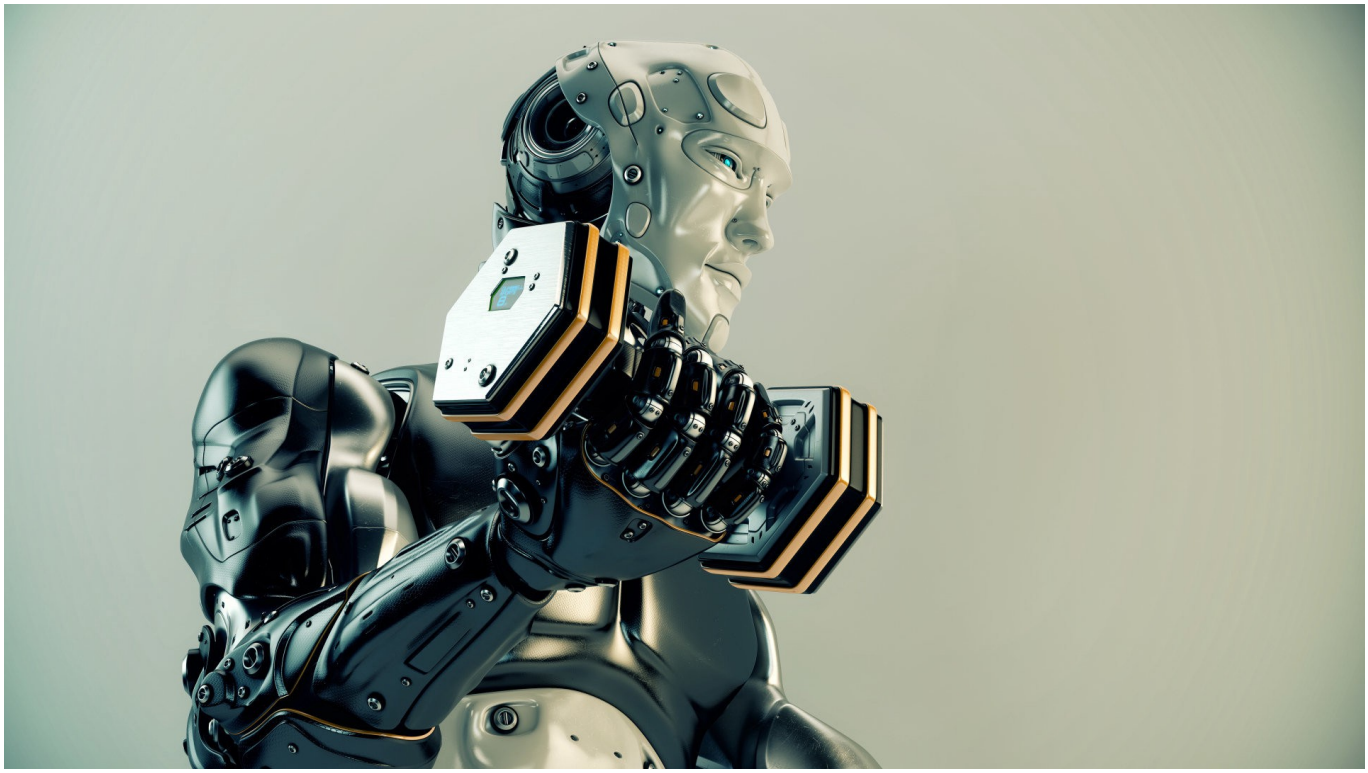
Tracking and sustaining the performance of predictive models

QuantumBlack Follow
Jul 30 · 6 min read

Jessica Fan, Product Manager and Musa Bilal, Machine Learning Engineer, QuantumBlack



Machine learning (ML) models have come to play a significant role in the business decisions of major companies around the world. Five years ago, many were asking whether these models could be harnessed to optimise corporate performance.

Today, the conversation has moved on. People are no longer asking whether deploying algorithms could theoretically drive organisational change — instead, the question now is how best practice can ensure that these models continue to deliver over time.

For all its benefits, the rise of ML has introduced a fresh range of risks for companies to grapple with. These include challenges in how to deliver ethical and explainable AI, but also the basic question of how — when data is playing an increasingly larger role in business decisions and our day-to-day lives — performance can be tracked and sustained.

We have written extensively around the broader themes to consider when harnessing ML models. In this article we will take a closer look at the build and maintenance process, exploring how organisations can create their own ML safety net to protecting against performance degradation — and unveil our latest tool, developed to deliver sustained, long-term ML success.

What causes performance degradation?

ML may be regarded as adaptable, consistently evolving technology, but we should always remember that as a tool it can sometimes prove fickle.

When a predictive model is created, it learns relationships between the input data and its predictive target. We may be confident of outputs when we use a dataset which has been repeatedly tested during development. However, how often in today's business world do variables remain static?

If the characteristics of data feeding the model change over time, the foundations of the relationships it is built on may become outdated. This can lead the model to start making unrealistic predictions –referred to as concept shift.

Additionally, if the nature of the input data become vastly different to the data used to train the model, the predictions can also become sub-par. This is known as covariate shift.

Performance degradation can also result from a positive feedback loop; the change in underlying data is the consequence of the positive change in behaviour the business is striving for, as a result of using the model's predictions. In short, the results of the model are so good, they change the status of the business, but the model does not take this change into account and so continues making predictions with previous performance in mind.

Why is this a challenge?

A common response to these challenges is to ask why we cannot simply retrain the model more frequently to keep it updated. Whilst retraining may address errant model behaviour, it's not an efficient or always affordable solution.

The alternative is to implement detection techniques into the analytics pipeline which can identify a 'true change' event rapidly and efficiently. There are countless ways of detecting concept or covariate shift, but how do we select the techniques which are suited to a particular degradation problem?

Moreover, the definition of "true change" is often non-trivial, and its definition will affect how often users are alerted about performance degradation. A model governing garden sprinklers may be able to afford a slight shift in timings if the result is a three-minute delay in plants being watered — it is an altogether different matter when a model is overseeing fire safety sprinklers.

And when we have confirmed the type of degradation to monitor for and how to detect it — how do we fix it?

Our approach to solving the problem

The latest internal tool developed by QuantumBlack, was designed to manage the performance of models we develop for clients. Our data scientists are using it across our client projects, during development to track the versions of models they build, and beyond project completion to detect performance degradations.

One of the tool's major features is the range of techniques it offers to detect covariate and concept shift. To help select the right one, our tool provides a grid search module that analyses previous cases of model performance deterioration, compare them to the model in question, then suggest the most appropriate technique configuration to detect degradation points as accurately and quickly as possible.

The insights are then shown in a dashboard, which can be used to proactively monitor model performance. This feature makes our tool a valuable feature for QuantumBlack clients, as it supports their ongoing efforts to spot model misbehaviour and correct it soon as possible.

The QuantumBlack team is currently experimenting with an auto-fix module to address degradation from covariate shift, which circumvents the problem of lacking readily available 'ground truth' data to retrain the model. The module implements a well-established approach that re-weights samples in the original data using similar, fresh data samples, and retraining the model on this adjusted set.

Regardless of the technical solution used, the following questions should be asked when determining how to monitor your models:

What methods / metrics should I use?

There are various techniques for detecting covariate shift, such as hypothesis tests, Bayesian change point detection, model based covariate shift detection. We recommend running a selection of these on historical data where a known change has occurred, to decide which technique identifies the change most accurately and quickly. If ground truths are readily available, you can extend the evaluation metrics used during training, and track this value over time with tools such as RMSE for regression problems, or precision rate for classification problems.

Is there a trade-off between false positives and false negatives?

What is the cost associated with a false alert versus failing to alert over an actual degradation? This can help decide whether the monitoring approach should tilt one way or the other, depending on how it could impact your business.

How often should monitoring be carried out?

Decide how often to run the monitoring process alongside normal pipeline runs. Our recommendation is to monitor whenever the data pipeline is updated or ingests new data.

How granular should measurements of performance shift be?

Decide aggregation levels of data. This is important when data is at a low level of granularity, as some techniques would fail to detect a change, or raise too many false alarms.

What threshold should be used for warning alerts?

Thresholds for different monitoring approaches should be decided. They can be set using historical performance of a particular metric, or simply by domain expertise.

How often should the monitoring approach be reviewed and updated?

Changes in data or business processes can invalidate monitoring techniques, so it is important to determine how frequently your approach should be reviewed and revised. However, this will be unique to each model and depend on its use and where it is sourcing data from. Decide how often a review is required once the monitoring approach has been finalised.

Supporting continued, consistent performance gains

Machine learning is often mistakenly viewed as a panacea to business problems, but while the technology may help alleviate some organisational issues, it also introduces its own set of challenges. We developed our tool to address these, but your organisation can also start creating its own ML safety net by answering the questions above.

As ML has become more widely adopted, our focus must shift from asking ‘is this possible’ to ‘is this best practice’. We hope this guide and our internal tool will help

ensure the technology continues to deliver its most crucial output — consistent, marginal performance and efficiency gains.

Machine Learning Artificial Intelligence Data Science Predictive Modeling