Broadstone

Effective Monitoring for Machine Learning Models in Consumer Lending

March 2025

Introduction

This whitepaper discusses some of the key considerations for the successful governance of machine learning models, particularly in relation to model monitoring.

These insights have been gained from our experience across the industry, and we discuss key elements of monitoring, potential pitfalls and important metrics.

Whitepaper Contents

Disclaimer

Broadstone Regulatory & Risk Advisory Limited is a company registered in England and Wales with Companies House number 04663795 with its registered office at 100 Wood Street, London EC2V 7AN. The company uses the trading name Broadstone, which is a trademark owned by Broadstone Corporate Benefits Limited and used by companies in the Broadstone group.

Rise of the Robots

The adoption of machine learning modelling techniques in the financial services sector continues to increase.

Ever expanding sources of data together with an ongoing drive for better, more accurate, lending decisions are leading more lenders to consider the benefits that AI and machine learning can bring.

But there is some reticence too. This remains a relatively new technology within the sector and this adoption is taking place in an environment of intense regulatory and compliance scrutiny. Lenders need to be able to defend and justify their lending decisions, ensuring they can be shown as aligning to the principles of the Consumer Duty and delivering fair customer outcomes.

The UK regulator is clearly focussed on the use of AI in lending; the FCA's current 'AI Lab' initiative is aimed at ensuring "safe and responsible use of AI in UK financial markets". So, lenders are cognisant of the very real benefits that this technology can bring, while needing to ensure that it is developed and deployed in a way that will not cause problems in the future.

This has shaped the initial use cases for AI and machine learning in lending. Efficiency tools driven by large language models (LLMs), such as internal applications for document summarisation or parsing text fields, or external customer-facing chatbots, are relatively common. These are not decisioning processes; they are facilitators for customer interaction or process completion.

In a similar vein, machine learning models have seen extensive use at 'non-exclusionary' decision points, such as pricing and fraud. While a customer may have an interest rate offered, or be referred for additional diligence checks, they are not being excluded from any product or service at these points.

But now lenders are looking to use AI and machine learning to support the modelling and decisioning processes throughout the business, including within the core underwriting risk models and critical regulatory processes. Doing this requires important stakeholders to be comfortable and confident in the models, and key to achieving that is effective governance and monitoring.

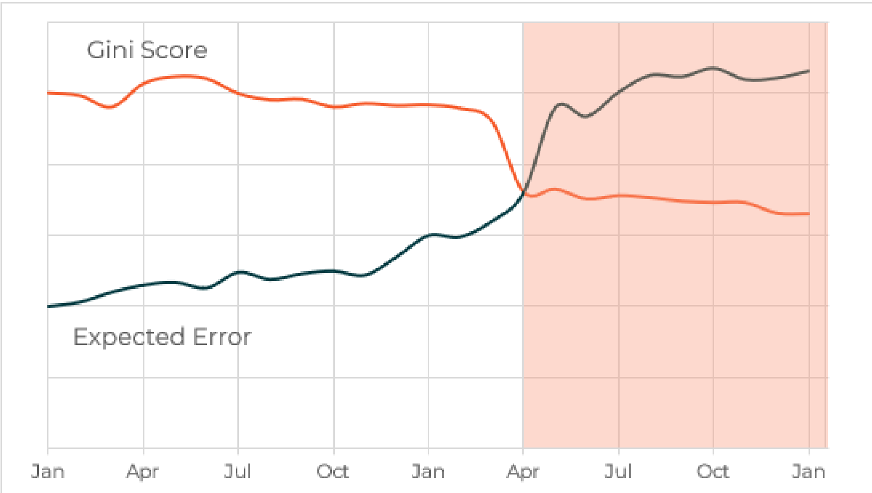
Fear of the Unfamiliar

The apprehension surrounding machine learning models stems from both internal and external sources.

For lenders, the executive team needs to be comfortable with the risks and benefits of implementing a new and often unproven technology. Supporting this, the compliance team needs to ensure that outcomes are fair, there is no discrimination or unacceptable biases and that data is being used appropriately and legally. In the absence of a deep analytical understanding themselves, these teams need to be confident that controls are in place to provide the required oversight. Key models, particularly those linked to regulatory processes, will fall within scope of the internal and external audit functions.

Model owners need to be able to effectively communicate the key features and drawbacks to these stakeholders, tailoring that information to the audience. Compliance may be most interested in proving that the model doesn't discriminate against those with protected characteristics, while auditors may need to be assured of the model accuracy and stability.

Sub-population Monitoring



Monitoring model performance and expected error on key sub-populations, such as vulnerable customers, helps ensure no unacceptable bias is present – essential for buy-in from compliance and legal teams.

Externally, there is expected to be increased scrutiny from the regulator on the use of ML models as their prevalence in lending grows. The Consumer Duty requires firms to demonstrate that their lending decisions result in good outcomes for customers; this must be evidenced clearly, regardless of the modelling approach used. Upcoming regulation will shape the expectations and requirements for lenders using AI technology.

The upshot of this is that to survive and thrive within the realm of AI and machine learning driven lending, strong governance and controls will be essential. While Model Risk Management is an entire topic in itself, for the purposes of this paper we will focus on one particular, analytically focussed, component of this: model monitoring.

Marking Your Own Homework

Model monitoring has always been an important part of using models effectively in a lending business, but this is even more critical for machine learning models.

Monitoring is an essential component in satisfying compliance, audit and regulatory needs, proving confidence that models are functional, sensible and aligned with business expectations.

Much has been written about the lack of transparency in machine learning models – 'black box' algorithms which take in customer data and then calculate a decision or score without giving clear insights into how it was obtained. There have however, been significant improvements in this area, with techniques and tools available to support model 'explainability'. Monitoring is a key component of this, giving lenders a peak 'under the hood' to ensure that the engine is running smoothly.

Monitoring reports should be designed with analysis sections clearly aligned to the needs of the different stakeholders. This will facilitate understanding and buy-in from the business, establishing clear responsibilities for sign-off of the various model aspects. Reports and tools should also be designed to hold up to scrutiny from regulators and to fit effectively within the framework of controls-based audit.

Often, the more sophisticated a model is the less immediate transparency there is in the underlying algorithms. Consequently, the monitoring needs to do more heavy-lifting to provide the clarity and security required. However, good monitoring can help overcome the inherent challenges present throughout the development and deployment lifecycle.

Taking Off the Training Wheels

Model monitoring can be particularly useful shortly after launching a new model - a fact often overlooked by lenders who focus on monitoring solely to identify model deterioration.

The technical complexity of building the pipeline infrastructure required to pass potentially hundreds of variables to the model in the correct format can easily result in small changes which can significantly affect performance. While this should theoretically be identified during model implementation testing, this is often carried out in an artificial UAT or test environment; the monitoring process will evaluate the true model results in the live environment.

Some real-world examples: in one case, a minimum instead of a maximum in the data ingestion code resulted in one key feature always taking a zero value in the live process. This was not spotted for a substantial period and led to several cohorts of underpriced lending with significant financial impact.

A second example involved a difference in credit bureau settings between the development / test data and the live environment; feature profiles were markedly different in some cases and the machine learning model performed poorly. The good news in this case is that while this was not detected in implementation testing, it was identified by a monitoring report being run on a weekly basis during the period immediately after the model's launch.

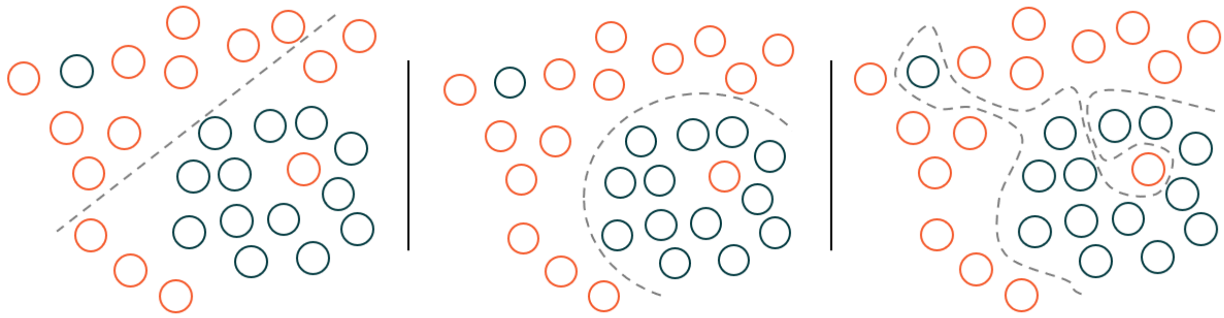
"Monitoring is a guard against model development errors or oversights that either would not occur with traditional models or would be much more trivial to spot."

Even when there are no issues with the infrastructure or implementation itself, machine learning models may not perform as expected post-launch. Monitoring is a guard against model development errors or oversights that either would not occur with traditional models or would be much more trivial to spot.

This is especially true when a model has been overfitted and does not generalise as well as expected to the through-the-door population. Overfitting occurs when the model overly reflects quirks or anomalies in the training data which don't occur outside of that data set. As a result, when it tries to apply the 'patterns' learned to a real-world situation the model makes incorrect assumptions.

There is a substantial risk of overfitting in machine learning models, which often go through thousands of iterations, compared to standard models where a relatively small group of 'sensible' variables is used. Automated machine learning methodologies lack the human intuition required to recognise what is likely to be noise in the sample versus a genuine trend.

While there are analytical techniques that can be used during development to mitigate overfitting risk, it is still common to see a large delta between expected and actual performance once a model is live.



Underfit Model   Balanced Model   Overfit Model

Effective monitoring needs to be able to identify overfitting or other development oversights quickly, before the actual full outcome has been observed. This can be achieved through tracking key sub-populations, monitoring referral rates to manual queues (and the resulting outcomes) or looking at early indicators for the eventual modelled outcome. When this is the case, monitoring serves not only as a tool to identify model deterioration, but also the final, most important step in the model development validation phase.

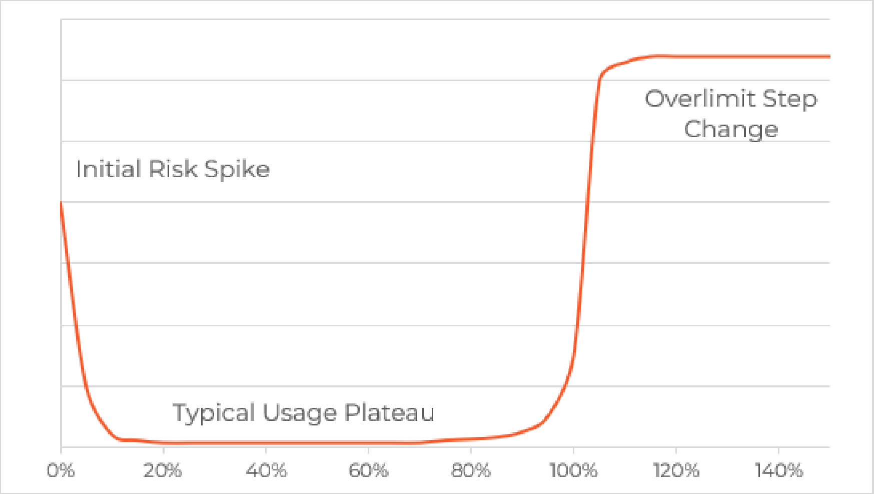
Managing the Ups and Downs

The longer-term monitoring considerations for machine learning and AI tools align closely with those of traditional models, but special considerations are required.

ML models can bring unique challenges or in some cases exacerbate potential issues. One important example of this is the stability of the input data. Traditional models use only a limited number of features, with a clearly understood (often linear) relationship between variable shifts and model impact. Monitoring reports will typically check the distributions of these features over time, assessing the stability of the modelled population, and triggering a retune or full redevelopment when significant shifts are detected.

The challenge with machine learning models is that they can react in an unexpected manner under relatively minor changes in the underlying data. The strength of machine learning models in capturing non-linear trends and complex interactions also means that the responses of the models to shifts in data are no longer always intuitive.

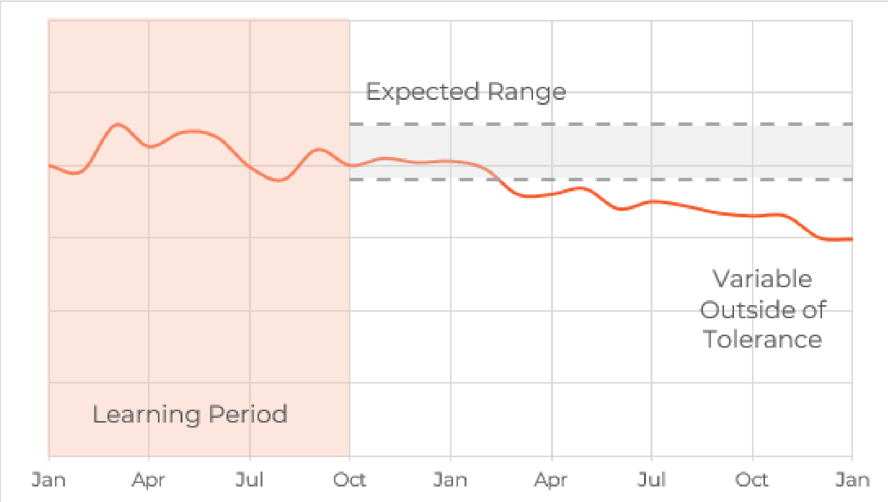
Credit Limit Utilisation vs Default Risk



Credit limit utilisation frequently has a non-linear trend which is often not captured in 'traditional' lending models but can be effectively utilised by machine learning models. However, this increases monitoring complexity as small shifts can result in a huge difference in risk.

As each model may use hundreds of variables, it is usually difficult or impossible for the model owner to assess them individually. Automated methodologies are required to analyse each feature, assess the potential impact of any profile shifts and flag potential issues. For example, the typical variance of each feature could be automatically established over an initial period, with deviation of this outside a specified tolerance being used on an ongoing basis to trigger a warning for manual review. Analysis can then assess whether the change represents a model risk, with 'false positives' automatically feeding back to increase the tolerance level.

Automated Anomaly Detection



An initial learning period can be used to establish 'typical' variance of a given variable and cyclical trends. This can then be used to determine an acceptable range for variable observations, where falling outside of this range results in a manual review.

Remaining on the topic of stability, we need to consider how the model responds to extrapolation. A relatively small shift in a variable can cause it to fall outside of the observed values used to train the model. While traditional algorithms will often be able to 'ride out' such changes with minimal or no loss of predictive power, machine learning models can be very sensitive to previously unseen values.

Once a change is identified, it is important to understand if it is significant enough to require remedial action or even a full redevelopment. In a standard regression scorecard, each feature plays enough of a role that a shift in any of them could compromise the model. Within machine learning models there are far more features with highly varying sensitivities; a given change could cause the model to degrade very suddenly, or conversely may have no impact at all. Additionally, the sheer number of variables in play increases the odds that a shift is observed in at least one of them, so this scenario is likely to arise much sooner than in a traditional score.

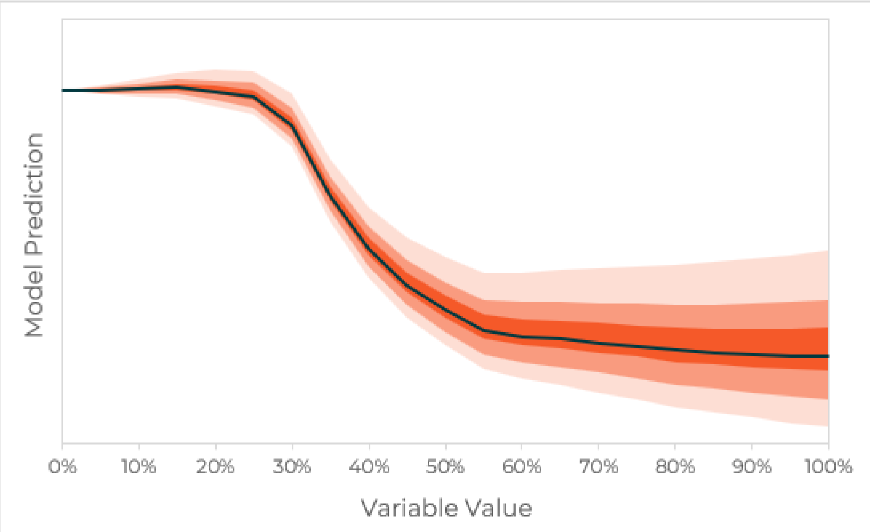
"A univariate view alone is no longer sufficient to understand model performance."

As well as looking at individual variables, monitoring must also consider variable interactions. Machine learning models are excellent at capturing complex relationships between variables. A variable could be relatively weak on its own, but combined with another becomes a strong predictor with a large impact on the model.

Due to this, a univariate view alone is no longer sufficient to understand model performance – in neural networks for example, the variables are combined and recombined together in intricate, non-linear ways. A relatively minor shift in two or more seemingly unrelated variables could cause a very large shift in the predictions generated by the model.

Effectively analysing these complexities can be facilitated by using more advanced visualisation techniques. Individual conditional expectation curves and partial dependence plots can help the business understand how individual and pairs of variables are functioning respectively. Incorporating these within the monitoring framework ensures that the business has the insight required to understand the model performance as the data landscape evolves.

Conditional Expectation Curve



Individual conditional expectation (ICE) plots show for individual instances how the model prediction changes as the variable changes but other datapoints are kept the same. Grouping individual lines into percentiles visualises changing variance through the variables range.

The final aspect of model monitoring is the one which often receives the most focus: identifying model deterioration. Lenders rightly look to the monitoring process to establish the divergence of the model from a theoretical 'optimum' and use this as a trigger for model redevelopment. Some machine learning models dynamically update on a 'self-learning' basis, but the use of these remains understandably rare in a regulated environment. This means that lenders still need to track model deterioration and understand when the current tools are no longer fit for purpose.

Machine learning is particularly helpful here; quick 'challenger' models can be built on the latest data to give a clear measure of what performance may be achievable. When the performance delta between this and the incumbent model is sufficiently large (i.e. big enough to support the associated business case), it provides the trigger for redevelopment.

As ever, there are caveats; these test models will usually inherit the data pipeline of the current model, so if there has been any substantial data revolution in the interim, such as the availability of brand new data sources, then potential benefit here may not be measured.

Charting the Way Forward

There are huge potential benefits from bringing AI and machine learning to bear within financial services.

Using more data to make better decisions more quicky is clearly an appealing goal, not just for lenders but also for the regulators, who want to see new technologies driving better customer outcomes. Because of this, we expect the usage of machine learning models to continue to grow rapidly in the industry, where benefits are already being evidenced in every part of the credit lifecycle, from underwriting to collections.

As this usage grows however, both the industry and the regulators are becoming increasingly aware of the significant risk posed by over-sold and under-monitored models. Firms need to take positive steps to ensure defensible and high-quality decisions are being made and to limit issues arising from poorly functioning or accidentally non-compliant models.

Fortunately, much of this risk can be managed and mitigated with the right governance and control framework, of which analytical model monitoring is a key part. Machine learning brings additional considerations and the need for more advanced approaches, but many of the fundamentals of model monitoring remain the same.

At Broadstone we have helped our clients build, deploy and monitor a wide range of model types. Whether you are in need of monitoring, or have broader development or model governance requirements, Broadstone are here to help you on your machine learning journey.

Broadstone

Find out more

For more information on how Broadstone can support your lending business, contact our credit risk experts:

Natasha Conradi
Senior Consultant
natasha.conradi@broadstone.co.uk

Paul Matthews
Senior Director
paul.matthews@broadstone.co.uk

Contact

020 3869 6900
corporate@broadstone.co.uk
broadstone.co.uk

@Broadstone_Ltd

in **company/thebroadstonegroup**

Birmingham / Bristol / Edinburgh / Glasgow / Liverpool
London / Manchester / Redditch / Sheffield / York