Good Practice Lending Guide

RM09 Model Monitoring Policy

September 2025

Disclaimer

This Guide is provided purely for informational purposes, has been prepared for general use only, and does not constitute legal, financial or other professional advice.

All information contained in this Guide is based on the laws and regulations applicable to England and Wales and which are current as of the date of publication This guide is not maintained regularly, but we will endeavour to update it when relevant laws or regulations are amended, varied, or supplemented. We aim to review the Guide annually to reflect any relevant legal and/or regulatory changes.

Fair4All Finance Limited make no representations or warranties of any kind, express or implied, about the accuracy, completeness, suitability, or reliability of the information contained herein. Fair4All Finance Limited shall not be liable for any loss or damage arising from the use of, or reliance on, this Guide. This Guide does not create an advisor-client relationship between you and Fair4All Finance Limited.

You are advised to consult with suitably qualified legal, financial or professional advisors to obtain advice tailored to your specific circumstances. You should not rely on the content of this Guide and any reliance on any information provided in this Guide is done at your own risk.

By accessing and using this Guide, you acknowledge and agree to the terms of this disclaimer.

This Guide must not be amended, copied, reproduced, distributed or passed on at any time without the prior written consent of Fair4All Finance Limited.

Note on legal and regulatory updates

This note should be read alongside the Disclaimer at the start of this Guide.

Laws and regulations can change. While this Guide is reviewed periodically, it may not always capture the latest developments.

For the most up to date information, please refer to the relevant official sources, including (but not limited to):

- **Financial Conduct Authority (FCA)**
  - FCA Handbook
- **UK Government**
  - Government policy papers and consultations
  - Credit union common bond review
- **Bank of England / Prudential Regulation Authority (PRA)**
  - PRA Rulebook for Credit Unions
  - PRA Supervisory Statements for credit unions
  - Policies relating to credit unions
  - Annual Assessment of the Credit Union Sector
  - Regulatory expectations
- **UK Legislation**
  - The Money Laundering, Terrorist Financing and Transfer of Funds (Information on the Payer) Regulations 2017 – (see RM6)
  - UK's Data Protection – (see RM 12)
- **Financial Reporting**
  - Financial Reporting Standards (FRS) – (see RM14)

Contents

1 Introduction 6

1.1 Which organisations is this document appropriate for? 6

1.2 Why is model monitoring important? 6

1.3 Purpose of this document 6

1.4 Scope and structure 6

2 Key principles of performance monitoring 9

2.1 Discrimination 9

2.2 Accuracy 10

2.3 Stability 10

3 Other considerations 11

3.1 Frequency of monitoring 11

3.2 Frequency of observation points 11

3.3 Exclusions 11

3.4 Level of assessment 12

3.5 Early performance emergence 12

3.6 Characteristic-level monitoring 13

3.7 Other business objectives 13

3.8 Benchmarking and tolerance 13

4 Governance 15

5 Worked examples 16

5.1 Worked example 1 – model discrimination 16

5.2 Worked example 2 – model accuracy 17

5.3 Worked example 3 – model stability 18

5.4 Worked example 4 – scoring strategy 20

6 Appendix 25

6.1 Appendix A: Glossary of terms used in the model monitoring module 25

1 Introduction

1.1 Which organisations is this document appropriate for?

This document is intended to support organisations that develop and use mathematically derived predictive models to support business decision-making. In particular, organisations who develop and maintain their own credit scoring models that are used to automatically risk assess customers when they apply for a loan or other credit product.

For organisations who mostly rely on expert judgement to make lending decisions or who only use a "Vanilla" (generic) credit score provided by a credit reference agency, the material in this document is unlikely to be relevant to the running of their businesses.

Good practice for using and monitoring the generic scores provided by credit reference agencies, usually supplied as part of a credit report, is provided as an appendix in the Lending Policy (Credit Risk) component of the Guide.

1.2 Why is model monitoring important?

Mathematically derived predictive models are used by many (mainly larger) organisations across the financial services industry and have an impact on decision-making and key outputs in a growing number of business areas. For example, credit scoring models are the main tool used by most high-street lenders for assessing the default risk of new customers. It is therefore important to understand how models perform on an ongoing basis, to inform whether they remain fit for purpose and identify any performance issues and/or triggers for further investigation or remediation. This process of assessing models is referred to as "model monitoring."

1.3 Purpose of this document

This document is intended to serve as a guide to the principles and best practices that should be adopted when monitoring model performance. An overview of how credit scoring is developed and applied is provided in the Lending Policy (Credit Risk) component. Some of the themes also overlap with other areas of Lending, such as Risk Appetite, Model Risk Management and Governance; these will be signposted within the relevant sections throughout this guide.

1.4 Scope and structure

This guide is applicable to all UK-based lenders but is primarily intended for small to medium sized

lenders, such as credit unions, who may have limited experience previous of developing and using mathematically derived models within their organisation.

The range of model types covered is broad, and includes (but is not limited to) the following categories and sub-categories:

- **Operational (credit scoring) models** – used primarily for decisioning and strategy at all points of the customer journey, including:
  - **Application scorecards** – used primarily for decisions on new credit applications
  - **Fraud scorecards** – used to identify potentially fraudulent applications
  - **Affordability models** – used to determine whether a customer is expected to be able to afford the repayments of a credit obligation
  - **Behavioural scorecards** – used primarily for decisions and strategy on existing customer accounts
  - **Collections scorecards** – used to inform prioritisation of collections efforts on delinquent or impaired accounts
- **Regulatory models** – used to drive reported financial outputs, including:
  - **Impairment models** – used to drive outputs reported for impairment provisions
  - **Capital models** – used to drive regulatory capital holdings
- **Business planning models** – including forecasting and stress-testing, to inform growth strategies, pricing, etc.

Not all of the above will be applicable to all lenders. When establishing model monitoring policy, lenders should specify the range of model types in scope, aligning with company structure, risk management and governance.

For the purpose of this document and its intended audience, the primary focus is on operational models and application scorecards. These generally come in one of two forms.

- **Bespoke models** - these are developed using the lenders own customer data and tailored to their specific objectives
- **Generic models -** these are developed for a specific industry sector, using cross industry data. Well know examples of generic models are those provided by Credit Reference Agencies (CRAs). Examples include Experian's Delphi Score, Equifax's Risk Navigator Score and TransUnion's VantageScore. For these types of models, the models themselves are not supplied to lenders, only the scores that the models generate for each customer

In the latter case, some of the more granular monitoring approaches discussed in this document may not be viable or relevant given that the low-level detail about how model scores are generated are not provided by the CRAs. However, the key performance metrics that impact business performance can be monitored for all model types.

In the remainder of this document, Sections 2 and 3 describe the general principles of model monitoring. Examples of how these are applied in practice are then provided in Section 5. Details of the derivation of the most common statistical measures used for model monitoring are described in Appendix A.

2 Key principles of performance monitoring

Different model types vary in terms of their design and purpose, which gives rise to different ways of monitoring their performance. However, there are common principles that generally apply, regardless of model type.

Broadly speaking, performance monitoring requires an assessment of predicted outputs from the model in comparison with actual outcomes. A certain outcome window must elapse before actual outcomes are known, which means performance cannot be readily assessed for more recent observations.

Most assessments of model performance fall into the following three areas:

- **Discrimination** – how well the model ranks individual accounts/customers in terms of the target variable being predicted
- **Accuracy** – how well aligned the model estimates are on average compared to the target variable
- **Stability** – how much model outputs vary over time

Each of these aspects is covered in more detail in the following sections.

2.1 Discrimination

In principle, the better a model discriminates or rank orders risk across a population of interest, the more successful it will be at meeting its business objectives.

This is especially important for operational scorecards, where score cut-offs are typically used to make credit decisions and hence the ability of the model to differentiate between higher and lower risk accounts/customers is key to meeting business objectives. In this context, the level of discrimination specifically around the cut-off points used for accept/reject decisions is key.

The measures used to assess discrimination may vary according to the type of target variable. For example:

- Binary target variables, such as default vs non-default, typically use measures such as the Gini coefficient.
- Continuous target variables, such as expected loss, typically use measures such as the coefficient of determination, R-squared.

Scorecard discrimination typically degrades over time for several reasons. This includes population shifts, changes in economic conditions, legislation, and consumer behaviour. In some circumstances this may be rectified by relatively minor updates to the model, such as a recalibration using the same variables. In other cases, a full rebuild is the best course of action. For example, if the current set of variables are no longer effective for the population of interest or additional data source(s) are available that were not originally presented to the model.

2.2 Accuracy

The accuracy of a model reflects how well it is calibrated to the target variable it is measuring. It is commonly assessed by comparing average predicted against actual outcomes. This may be evaluated at a portfolio-level, and/or for smaller segments or sub-populations of business interest.

Accuracy is important for most model types. For operational models, poor accuracy can easily lead to a significant under/overestimation of outcomes, even if the model discriminates risk well. In this scenario, it may be appropriate to recalibrate the model so that predictions are well aligned to actuals based on the latest available performance data.

2.3 Stability

It is useful to assess model stability over time, as it gives an indication of portfolio change and its impact on the distribution of model outputs. Unlike with discrimination and accuracy, stability monitoring only requires predicted and not actual outcomes, hence the most recent observation periods can be assessed.

In contrast with discrimination and accuracy monitoring, changes in the output distribution over time do not necessarily indicate a problem with the model. For example, the growth of a new product type within a portfolio could drive a distributional shift towards a certain score range. If the model is effective for the new product type, then such a shift is appropriate and good discrimination and accuracy will be retained.

Population stability is commonly monitored by tracking the Population Stability Index (PSI) over time, which summarises the consistency of the distribution of model outputs across quantiles.

3 Other considerations

3.1 Frequency of monitoring

The frequency of monitoring report production can differ according to several factors, including:

- **Risk management and governance requirements** – a certain frequency of performance assessment may be required to conform with internal risk management or policy, or audit/regulatory requirements. This should be proportionate to the size and systemic importance of the firm (eg, whether the firm has AIRB status for calculating capital requirements).
- **Product(s) to which the model applies** – for low volume, relationship-managed portfolios, it may be appropriate to monitor on a less frequent basis, if more regular tracking would yield insufficient additional volumes to drive a statistically meaningful change in performance. Other product types may also warrant less frequent monitoring regardless of volume, eg, products with less frequent repayment schedules.
- **Internal data production** – monitoring processes typically require snapshot views of the portfolio(s) to which the model is applied, together with any supplementary performance data, such as loss or write-off data used in LGD monitoring. The timing and frequency of any input datasets required may dictate what frequency of monitoring is practicable.

Most commonly, lenders tend to produce monitoring reports monthly, in line with input data conventions. However, a formal assessment of these reports by an appropriate committee may be less frequent, eg, quarterly.

3.2 Frequency of observation points

Monitoring reports are mainly constructed with observation periods on the horizontal axis (see Section 5 for examples), providing a trended view. The frequency of these points would normally fall in line with the data snapshot convention, most commonly monthly. However, there are instances where it may be appropriate to change this frequency, for example where monthly points yield insufficient volumes leading to volatile trends, it may be preferable to group observations into quarterly or even yearly points.

3.3 Exclusions

It may be appropriate to exclude certain records from the population of interest when preparing the data for monitoring purposes. This is situation-specific, but in principle the aim should be to ensure the selected data is relevant to the model assessment. As a starting point, consideration should be given to

aligning the exclusions used for monitoring with those used in the original model development or when the model was first deployed (for a generic model). Examples of appropriate exclusions could include:

- Accounts with missing or default scores
- Fraudulent applications
- Accounts already in default at point of observation[1]
- Accounts that are deemed inactive, closed, or have a low balance at point of observation[2]

3.4 Level of assessment

The modelling-level, ie, the level at which the modelling dataset is structured, and scores are produced, dictates the granularity of model outputs. This will usually be at the account or application-level for operational models, ranging through to coarser segment or even portfolio-level structures for some forecasting and business planning models.

The data is typically aggregated for monitoring purposes, with the highest level of assessment being the total portfolio-level. This view is usually important, and in some instances may be the only view required, for example, where low volumes preclude any more granular assessment. In many cases, lower-level breakdowns are also required, which could be separate model segments or specific sub-populations of interest, such as:

- Sub-products or brands within the portfolio
- Applicants with or without credit history (for application scorecards)
- New vs seasoned accounts (for behavioural scorecards)
- Accounts up-to-date vs in arrears (for behavioural scorecards)

3.5 Early performance emergence

A common challenge with both building models and monitoring them is that they often require a lengthy outcome window. For example, for an application scorecard, it is usually to use an outcome period of 12-18 months to observe their performance. This restricts the available period over which actual performance can be fully evaluated, meaning that the latest performance assessments relate to older observation points. Consequently, the feedback loop is delayed and significant changes in performance may not be realised until long after the model predictions were made.

One way of dealing with this restriction is to create alternative monitoring views based on shorter outcome periods. In doing so, it may also be appropriate to change the severity of an outcome definition to limit the impact on available volumes. For example, assessing the roll to 2 missed payments over 6

[1] Not applicable to application scorecard monitoring
[2] Not applicable to application scorecard monitoring

months may be a suitable proxy for roll to default in 12 months, and enables assessments to be made on observations as recent as 6 months prior as opposed to 12 months prior.

3.6 Characteristic-level monitoring

The performance measures considered in Section 2 relate to the model outputs or scores. Depending on the model construction and complexity, it can be beneficial to also assess performance of individual characteristics. This can help to identify which specific characteristics are responsible for any deterioration in model performance.

Accuracy of an individual characteristic can be monitored by assessing the alignment of predicted and actual values of the score or target variable across the attributes of the characteristic.

Stability can also be assessed at the characteristic-level, again using a metric such as PSI calculated across the attributes or quantiles of each characteristic. Any significant population shift on the model output, or score is likely to be driven by a shift in one or more of the characteristics, although again, this does not necessarily indicate an issue with the model.

3.7 Other business objectives

Much of the focus has been on the ability of the model to predict its intended outcome, such as predicting, at the time of application, the risk of a customer defaulting on any credit they are provided with. However, there may be other relevant measures from a broader business perspective, such as revenue, loss, or NPV. Such considerations may be incorporated as part of the model monitoring and may support ongoing strategies, eg, by understanding the trade-off of increasing accept rates[3] and default rates from lower score cut-offs.

3.8 Benchmarking and tolerance

For monitoring to be meaningful, there needs to be a sense of expected performance. The level of performance achieved in the model development[4] often serves as a suitable benchmark for ongoing performance. A certain degree of degradation can be expected over time, which may be tolerable to a point, with any further deterioration serving as a potential trigger for model remediation or rebuild.

Periodic assessments should also account for volatility caused by limited volumes of accounts or (especially) certain outcomes[5] contributing to individual data periods. This can be mitigated, for example, by grouping consecutive periods to boost the available volumes, or by assessing against an alternative

[3] And consequently, increased revenue.
[4] Strictly speaking, this should be based on an independent test set as opposed to the dataset used for training the model, which will often overstate model performance. Where a generic score is used, then analysis of the performance of the score should be established (and hence benchmarks set) using test samples of the lender's customer data prior to the model being used operationally.
[5] Eg 'bads', in the case of a probability of default (PD) model.

target definition that affords more instances compared to a rare outcome[6].

Taking account of these principles, a typical monitoring schema may prescribe tolerance thresholds for ongoing performance. Threshold breaches typically serve as a trigger for further investigation of the cause, which in some cases could dictate that the model needs to be remediated or rebuilt, depending on the severity and persistence of the observed deviance and the reasons identified.

Suitable thresholds can be derived in various ways, the choice of which is often dictated by the metric under consideration:

- **Judgemental thresholds** – it may be appropriate to set judgemental tolerance thresholds, aligned to levels that would be viewed as concerning by the business. For example, when monitoring the discrimination of an application scorecard, it may be deemed that a 10% relative drop in Gini performance poses a material concern that warrants further investigation.
- **'Rule of thumb' based thresholds** – tolerance limits for certain metrics may be dictated by common industry convention, eg stability monitoring often assumes a 10% deviation in PSI as a moderate trigger and a 25% deviation as a severe trigger.
- **Confidence intervals** – can be statistically derived, with wider tolerance bands resulting where volumes are lower at a particular observation point. For example, for monitoring the accuracy of an application scorecard, the expected default rate could be monitored with thresholds set above and below corresponding to a 90% confidence interval, with an assessment of whether the actual default rate falls within these boundaries.

[6] Eg assuming delinquency criteria less severe than the definition of default would result in more 'bads' and potentially a more meaningful assessment

4 Governance

As part of good governance, a lender's model monitoring should be reviewed on a periodic basis and any assumptions and tolerances updated as necessary.

Requirements for the production, stakeholder review and approval of model monitoring reports should be well defined, including any requirements for committee submissions.

A future document within this series will cover the topic of Governance in more detail.

Large organisations that use models widely across their organisation, or who fall within the remit of the PRAs model risk management principles for banks, will often have dedicated model governance committees to manage and monitor the suite of models that exist within their organisations.

Less model-centric organisations and those that are exempt from the PRA's model risk management principles (such as credit unions), will typically incorporate model governance into established risk or Board committees.

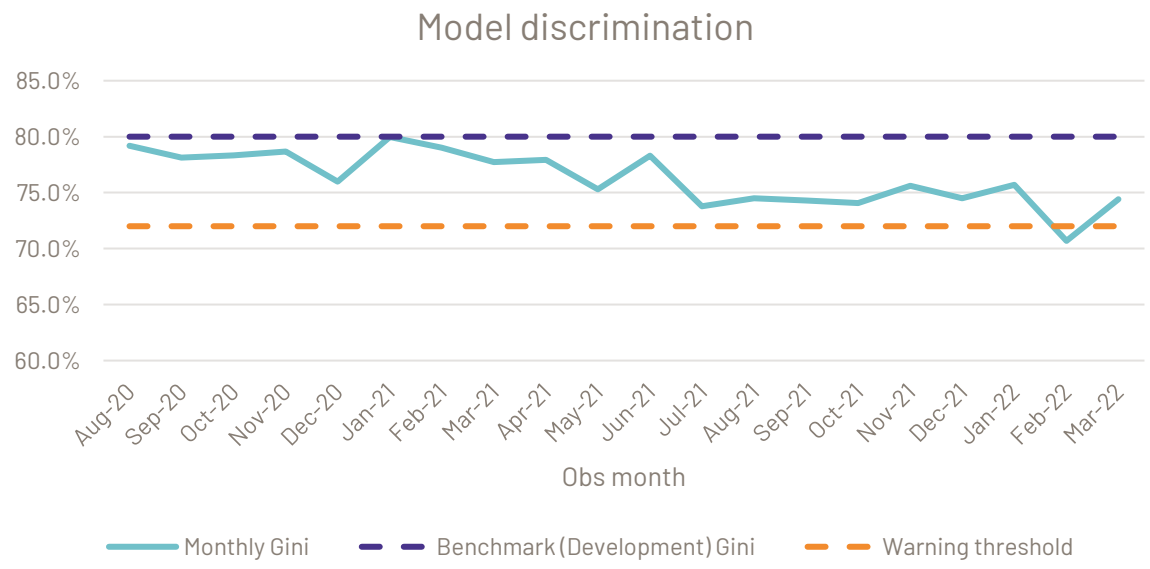
5 Worked examples

The following illustrative examples are intended to cover the monitoring principles outlined in Sections 2 and 3:

5.1 Worked example 1 – model discrimination

Example 1: Monitoring the ranking performance of a scorecard used in application or customer management decisioning.

The following chart shows model discrimination, measured by the Gini coefficient, over a varying observation window. The Gini measured on the original development data is shown as a constant benchmark, along with a warning threshold set at 10% (relative) below this. A 12-month outcome window is required to observe defaults, hence the latest observation point shown is 12 months prior to the latest available data, ie the March-22 data point represents the scores calculated in March-22 and their subsequent performance in March-23.



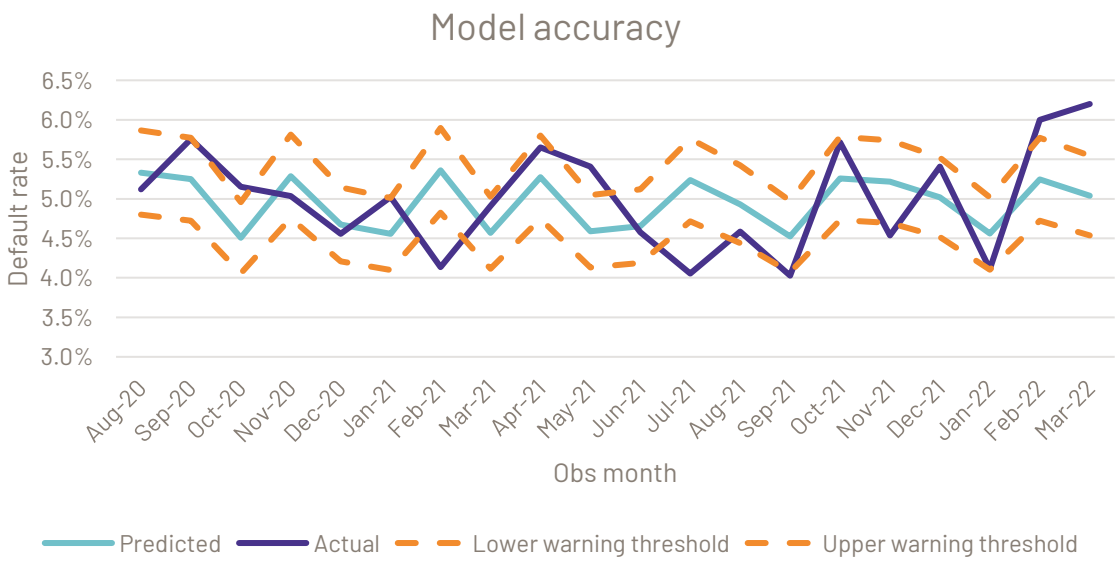
There is some evidence of a slight degradation in performance over time, as would typically be expected for a scorecard. A minor tolerance breach is observed in the penultimate observation month, which appears to self-correct and may be down to month-on-month volatility, depending on the volume of data

underpinning each monthly calculation. Based on the latest observed performance, there does not appear to be an immediate cause for concern, although a continued diminishing trend leading to multiple consecutive tolerance breaches would warrant further investigation, that could potentially support the case for replacing the model with a new, better performing, version.

5.2 Worked example 2 – model accuracy

Example 2: Monitoring the calibration accuracy of a model predicting the probability of default in the next 12 months.

The following chart compares average predicted default rates against those observed within 12 months post-observation. The lower and upper warning thresholds shown are set respectively at 10% (relative) below and above the predicted values.



There is greater volatility in the actual default rates than the model predictions, possibly a result of relatively low monthly volumes, leading to occasional isolated breaches of the warning thresholds. There is no evidence of a consistent over or underprediction, with the average default rate being fairly stable over time.

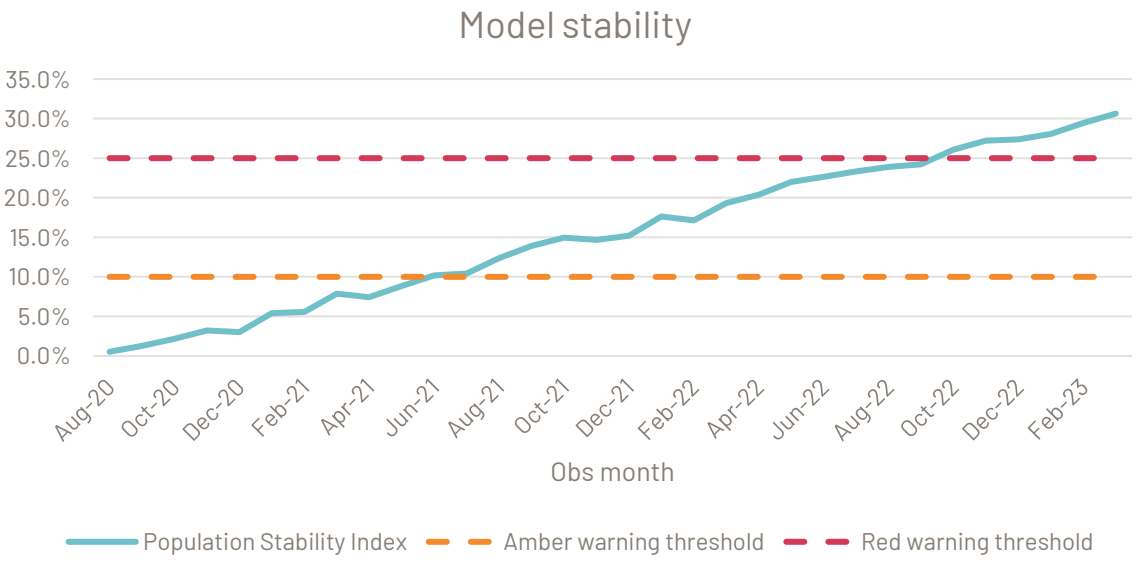
However, the two most recent observations show an underprediction that breaches the upper warning threshold in consecutive months. This may turn out to be within normal volatility although could be the start of an increasing trend, hence warrants further investigation. A useful assessment could be to

assess the trend in default rates over a shorter outcome window[7], to see if this is expected to self-correct or if default rates are projected to continue increasing in the subsequent observation months.

5.3 Worked example 3 – model stability

Example 3a: Monitoring the ongoing stability of model scores.

The following chart shows model stability, measured by the Population Stability Index, comparing monthly observation points to the original development data. The amber and red warning thresholds are set at 10% and 25% respectively, based on standard industry benchmarks of what constitutes moderate or more severe distributional changes. Unlike for the previous examples, the very latest available observation point is shown, as there is no outcome window required to observe subsequent performance.



There is evidence of a gradual population shift over time, with the red threshold being breached continually over the most recent observation months. There could be multiple reasons for this, including changes in marketing or acceptance strategy or other characteristics of the customer population, which could be detected via characteristic stability analysis[8].

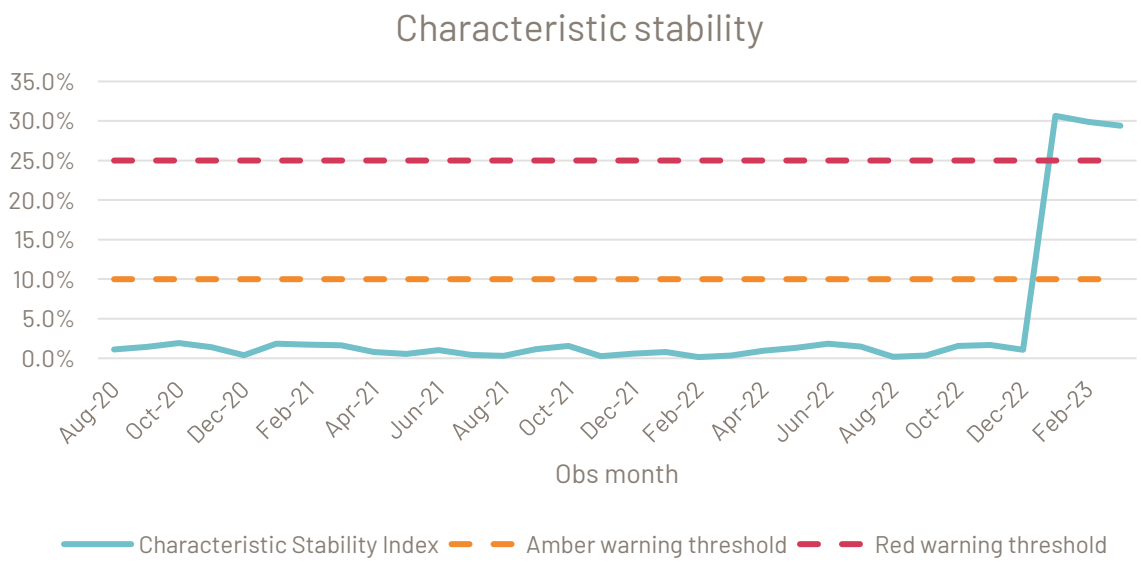
As noted in Section **Error! Reference source not found.**, a significant change in score distribution does not necessarily indicate an issue with the model, providing that performance is adequate on the new subpopulations. If, for example, the discrimination of this model[9] had remained consistent over a

[7] As described in Section 3.5
[8] See Worked Example 3b
[9] See Worked Example 1

sustained period of increasing PSI, this would support that the model remains fit for purpose despite a

Example 3b: Monitoring the ongoing stability of individual scorecard characteristics.

The following chart shows characteristic stability, again measured by the Population Stability Index, this time focusing on the range of values for a single model characteristic.



In this example, the distribution across the range of this characteristic is consistent for most of the time series until the last three points, where a step change is seen. Such a sudden shift could be due to a significant change in lending policy, such that certain attributes become much more or less frequent in the population or could potentially indicate a data issue.

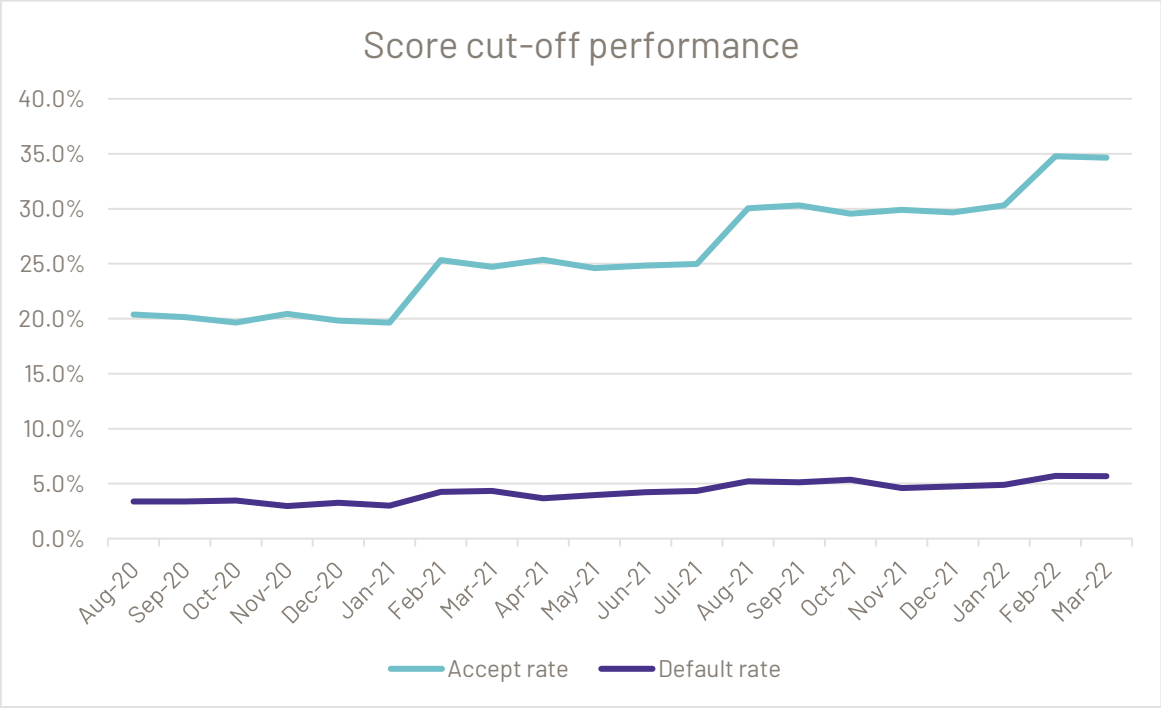
Immediate investigation is warranted, although again this may not necessarily indicate an issue if either the effect on the overall score is minimal[10] or the model is still effective despite the distributional shift.

[10] Eg, if this characteristic has a low contribution to the overall score

5.4 Worked example 4 – scoring strategy

Example 4a: Monitoring the impact of changing score cut-offs for application decisioning on accept rates and default rates.

The following chart shows simultaneous trends in accept and default rates over time for an application scorecard.

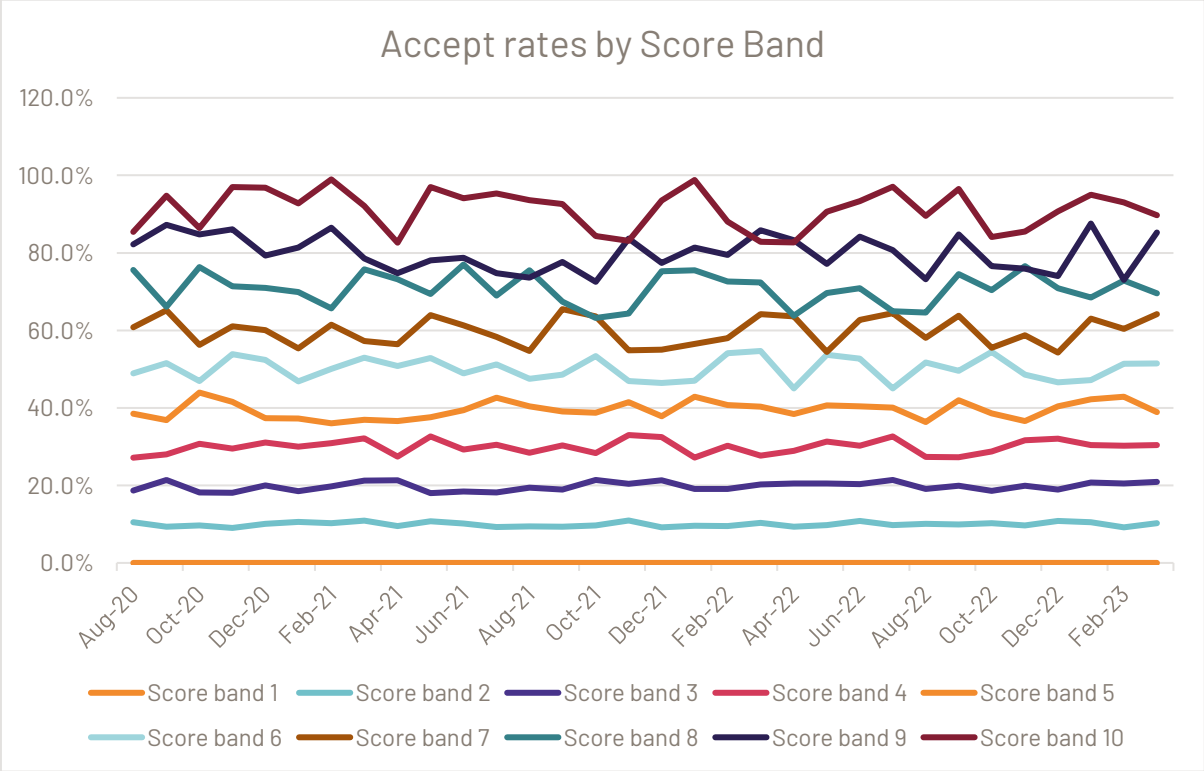


The step changes clearly seen for the accept rate trend correspond to points where the score cut-off was reduced. Naturally, this allows higher risk customers to be accepted, hence there are corresponding increases in default rate. Understanding this trade-off and where to set an appropriate cut-off is key to maximising revenues while remaining within risk appetite.

Note that this example only considers performance on all booked accounts and as such, does not reflect the marginal performance of accounts close to the cut-off, or indeed at any specific score region. The following examples provide a more granular view to enable more informed assessments.

Example 4b: Monitoring accept rates by score band.

The following chart shows accept rates over time, for pre-defined deciles of an application scorecard.

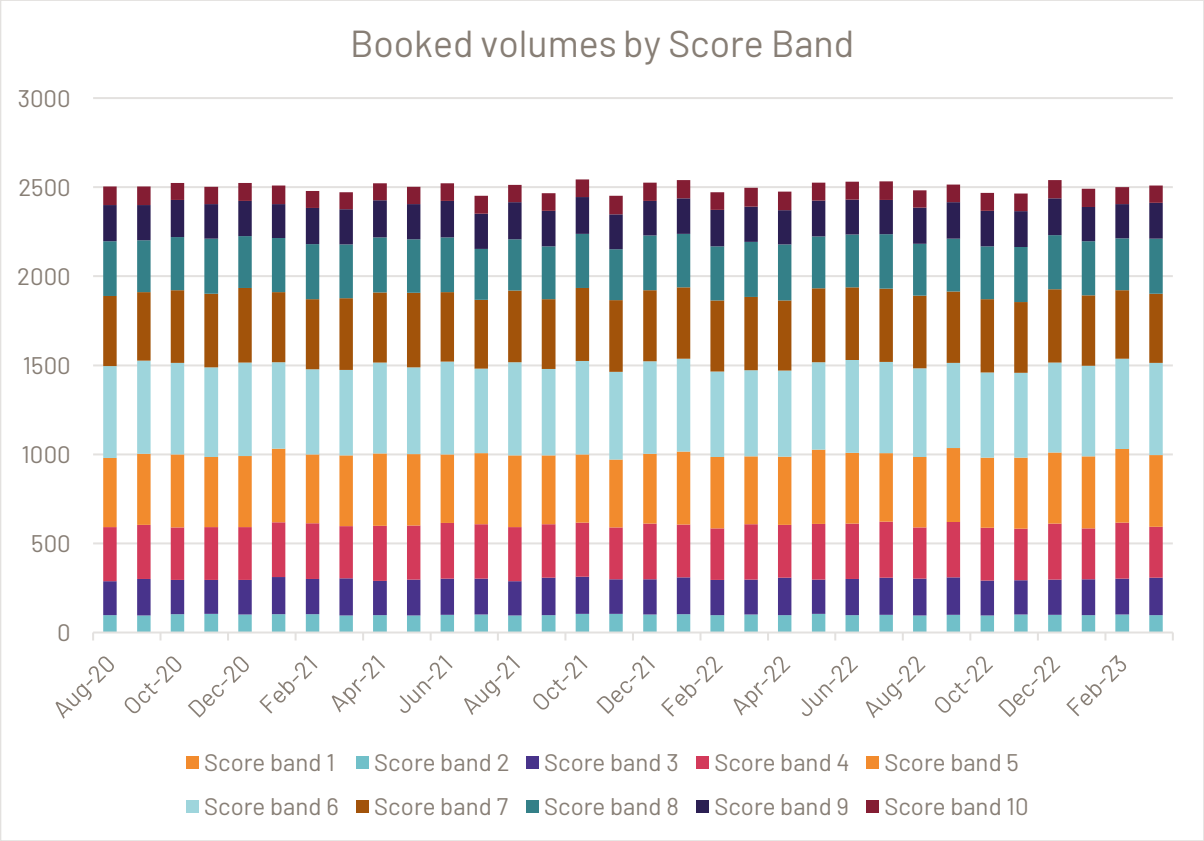


In this example, Score band 1 has a constant 0% accept rate due to meeting automatic decline criteria. As the score bands increase, gradually higher override rates for referrals drive higher accept rates. Towards the top end, all scores meet the score acceptance criteria and only other policy decline rules prevent 100% accept rates.

Such a report could be adapted to show override rates, which indicate how much the scorecard is relied upon in the overall decisioning process.

Example 4c: Monitoring booked volumes by score band.

The previous example focused on accept rates across the score range, which neglects the volume of applications falling within each score band. The following chart shows volumes of booked accounts[11] over time, again for pre-defined deciles of an application scorecard.



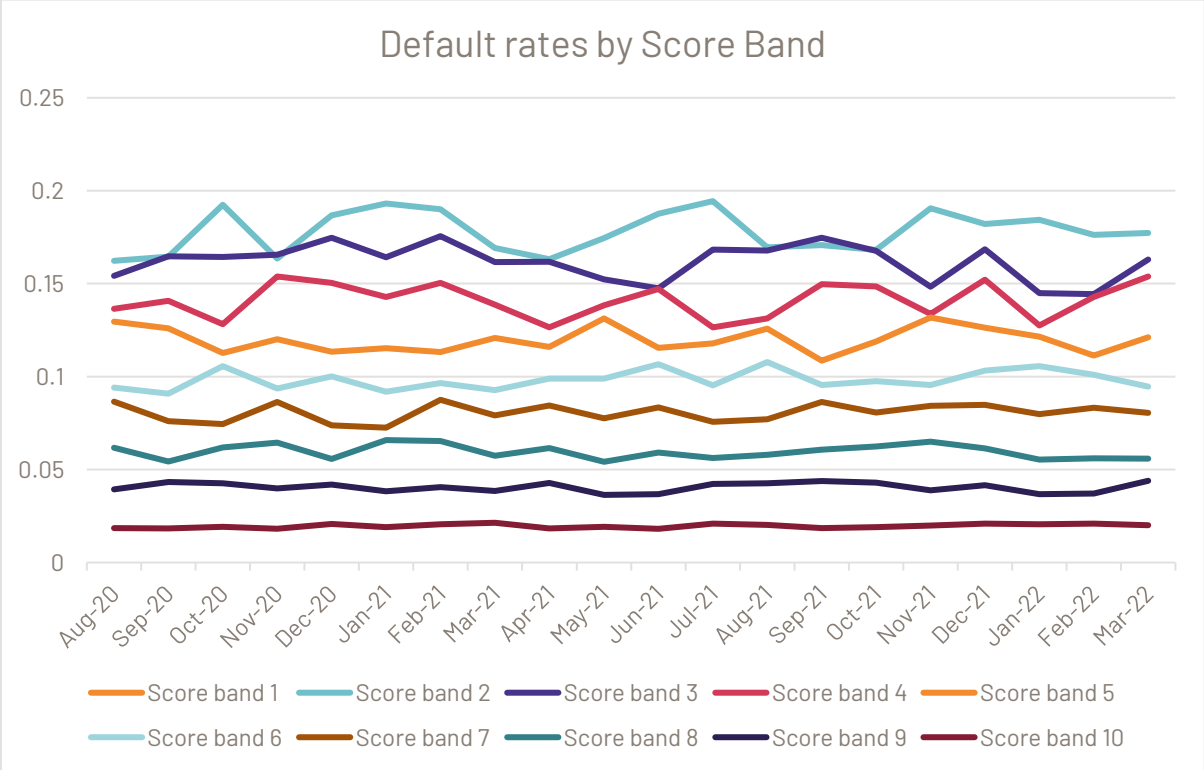
In this example there is a roughly normal distribution of scores across the range. Both volumes and proportions appear stable over time.

Such a chart could be considered in conjunction with the next example, which shows the default rate by score band.

[11] Note that the volume of booked accounts differs to the volume of accepted applications, as some offers are 'Not Taken Up' (NTU) by the applicant. In practice, it may be appropriate to monitor accepted and booked volumes separately

Example 4d: Monitoring default rates by score band.

The following chart shows the marginal default rate for each application score decile.



This view enables a more granular risk assessment, that could be used to inform ongoing lending strategy. For example, if a default rate above 15% was deemed too high to meet risk appetite, then potentially applications falling in Score band 3 or lower should be automatically declined going forward.

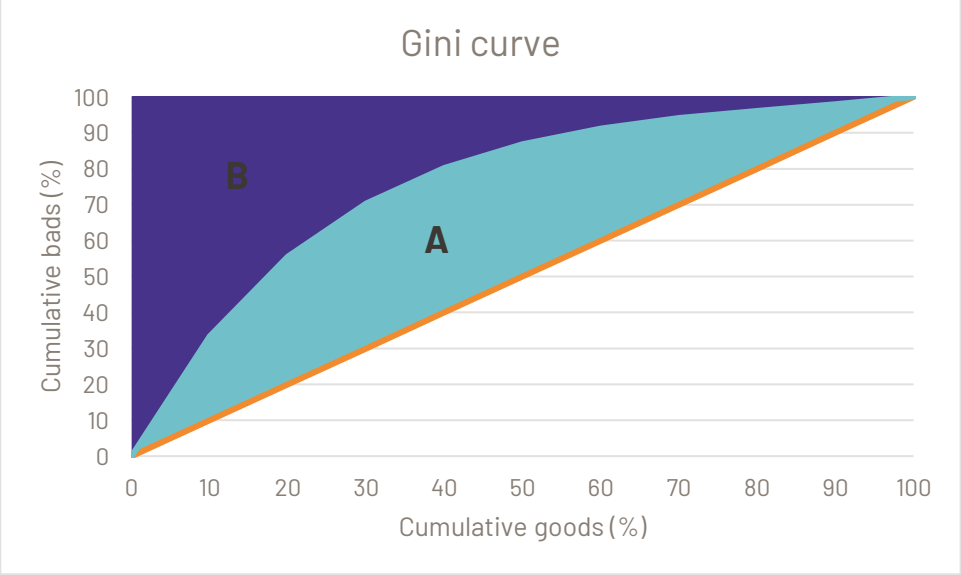
This chart also gives an alternative and more granular view of model discrimination. In this example, there is generally good separation across the score deciles over time. Increased crossover of adjacent bands would be a potential cause for concern, particularly if this happened in a score region close to the score cut-offs.

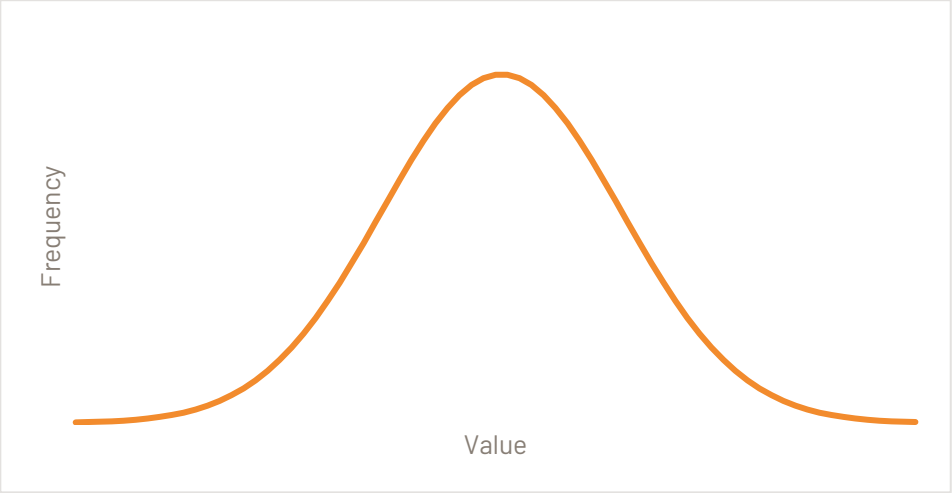
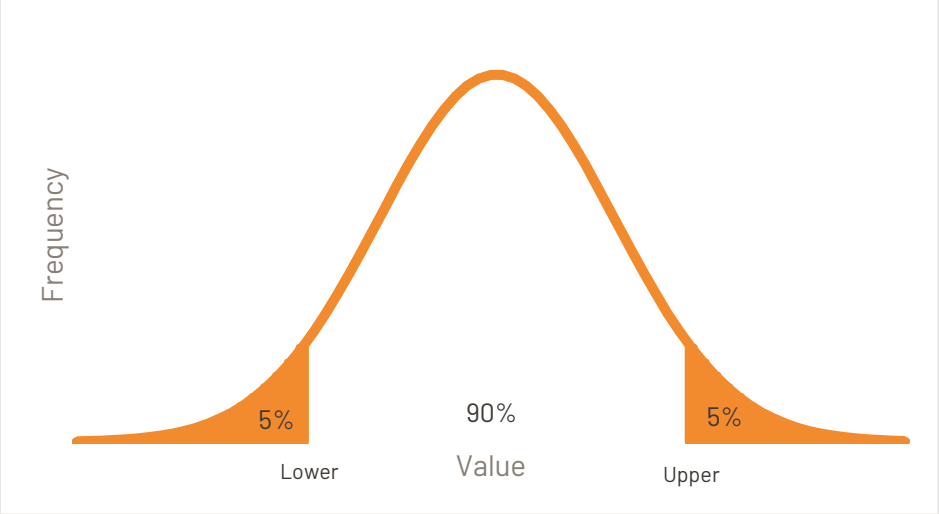
6 Appendix

6.1 Appendix A: Glossary of terms used in the model monitoring module

Please see the full Glossary for a comprehensive list of Credit Risk Management terms.

Term	Description
Predictive Model	A (predictive) model can be broadly defined as any process that uses past and current data to predict future behaviour of a specific outcome or to understand the current state of a system. Predictive modelling typically involves training and testing data and may involve one or more techniques, algorithms and tools including but not limited to: • Statistics • Mathematics • Machine Learning algorithms • Optimization techniques
Scorecard	A scorecard is a type of model that uses characteristics to calculate and assign a score to an individual, which represents their expected level of risk with respect to a pre-defined target variable.
PD	Probability of Default.
LGD	Loss Given Default.
Variable / Characteristic / Feature	Alternative terms for the required data fields used to calculate model outputs.
Target variable	The dependent variable on which a model is trained to predict the outcome of.
"Bads"	A term commonly used in credit risk modelling for accounts that move into a default

	state within a defined outcome period.
"Goods"	A term commonly used in credit risk modelling for accounts that remain not in default within a defined outcome period.
Gini coefficient	The Gini coefficient measures score discrimination (ie the ability of a scorecard to separate high risk from low), where the target variable is binary (eg default vs non-default). It is calculated by comparing the cumulative proportion of good and bad accounts throughout the scoring range: $$1 - \sum_i (B_i - B_{i-1})(G_i - G_{i-1})$$ Where: B_i = Cumulative bads at score i G_i = Cumulative goods at score i Or equivalently, as illustrated in the below chart: $$\frac{Area\ A}{Area\ A + Area\ B}$$  A random score with no discrimination would have a Gini of zero and a perfect score would have a Gini of 100%.

Normal distribution	A normal distribution describes data that is symmetrically distributed and follows a bell shape as shown below.  Many different types of data (approximately) follow a normal distribution and may types of statistical tests assume that the data being examined follows a normal distribution.
Confidence interval	A range of estimates, within which an unknown parameter is expected to lie, to a certain degree of confidence. For example, the 90% confidence interval of a normally distributed random variable is shown below:  The central (non-shaded) area under the curve represents the range of values within which we can be 90% confident that the true value lies.
R-squared / Coefficient of	A measure of the strength of a linear relationship between 2 variables, commonly used to assess the fit of a model designed to predict a continuous target variable.

determination	It is calculated as the square of the Pearson's Correlation Coefficient, r: $$r^2 = \frac{\left(n(\sum xy) - (\sum x)(\sum y)\right)^2}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$$ Where: x = individual predicted values of target variable y = individual actual values of target variable n = number of observations of predicted and actual values R-squared represents the percentage variation in the target variable (y) that is explained by the model prediction or score (x). The range is 0 to 1 (ie, 0% to 100% of the variation in y can be explained by x).
Population Stability Index (PSI)	PSI is an industry standard measure to identify significant shifts in distributions, commonly used to compare recent observation data to that used in the model development. It is calculated as follows: $\sum_i \left(\ln\left(\frac{obs\%_i}{dev\%_i}\right) \times (obs\%_i - dev\%_i) \right)$ Where: $obs\%_i$ = Proportion of observation population in band/category i $dev\%_i$ = Proportion of development population in band/category i This calculation can be applied to assess the stability scores or individual characteristics alike. In terms of what constitutes a significant change in distribution, this is somewhat subjective and various rules of thumb are applied in practice, such as 10% and 25% for moderate and severe changes respectively.[12]

[12] It is also worth noting that PSI is sensitive to the number of bands on which it is assessed; however, providing it is calculated consistently over time still provides a fair assessment of distributional trends