Risk Analytics

Machine Learning and Optimization
for Data-Driven Decision Making

Fernando S. Oliveira

Draft version — May 20, 2026

Chapter 9
Risk Scoring, Calibration, and Risk Decisions

Chapter 9

Risk Scoring, Calibration, and Risk Decisions

9.1 Introduction: From Prediction to Risk Decisions

Chapters 7 and 8 developed the predictive layer of risk analytics. Chapter 7 used classification trees to show how transparent rules can segment borrowers, customers, suppliers, assets, or projects into interpretable risk groups. Chapter 8 broadened the modelling toolkit to logistic regression, random forests, and gradient boosting, and showed how predicted event probabilities can support ranking and expected-loss analysis.

This chapter takes the next step. It asks how predictions become decisions.

A predicted probability is not yet a decision. A default probability of 8% does not by itself determine whether an account should be accepted, repriced, reviewed, monitored, or declined. A fraud probability does not by itself determine whether a transaction should be blocked. A supplier-disruption probability does not by itself determine whether inventory should be increased or a backup supplier activated. Predictions become useful only when they are embedded in a decision process.

That process requires a score scale, risk bands, thresholds, review capacity, calibration evidence, expected-loss interpretation, monitoring routines, and governance. The central message of the chapter is therefore simple: risk scoring is not model-output formatting. It is the design of a decision-support system.

A useful scoring system must answer four managerial questions. First, what does the score mean? Second, can the score be trusted for the intended use? Third, what action follows from each score range? Fourth, how will the organization monitor whether the score remains reliable after deployment?

The chapter completes the progression of Part III. Chapter 7 moved from data to interpretable risk segments. Chapter 8 moved from segments to probability prediction. Chapter 9 moves from prediction to calibrated, banded, monitored, and action-oriented risk decisions.

The chapter proceeds as follows. Section 9.2 explains how predicted probabilities become scores and risk bands. Section 9.3 discusses calibration and probability reliability. Section 9.4 connects thresholds, expected loss, and capacity-constrained decisions. Section 9.5 discusses model selection, interpretability, monitoring, and governance. Section 9.6 extends the credit-default case study from Chapters 7 and 8. Section 9.7 summarizes the managerial lessons.

9.2 Scores, Scorecards, and Risk Bands

This section explains how a model output becomes an operational risk tool. The model produces a predicted probability. The scoring system translates that probability into a scale that users can understand, communicate, monitor, and connect to decisions. The same underlying prediction can therefore appear in three forms: a probability, a score, and a risk band.

For case i, let Y_i be the event indicator, with $Y_i = 1$ if the adverse event occurs and $Y_i = 0$ otherwise. Let p_i be the model's predicted probability that $Y_i = 1$. The simplest possible risk score is the predicted probability itself:

$$S_i = p_i. \tag{9.1}$$

This probability score is directly interpretable. If the model is calibrated, a group of cases with p_i close to 0.10 should experience the event approximately 10% of the time.

Credit scoring often uses a transformed score rather than the probability itself. A common transformation is based on the odds of the adverse event. The odds associated with probability p_i are $p_i/(1 - p_i)$, and the log-odds are

$$\log\left(\frac{p_i}{1 - p_i}\right).$$

A points-based score can then be defined as

$$\text{Score}_i = A - B\log\left(\frac{p_i}{1 - p_i}\right). \tag{9.2}$$

With this convention, higher scores mean lower estimated risk.

The two constants in Equation (9.2) have different roles. The constant B controls the spacing of the score scale. It determines how many score points correspond to a given change in the odds of default. The constant A anchors the score scale. It determines the score assigned to a chosen reference probability.

The choice of B is a communication convention. Suppose the analyst wants a 20-point increase in score to mean that the odds of default have been halved. If one case has odds twice as large as another, the difference in log-odds is $\log 2$. Because Equation (9.2) multiplies log-odds by B, the score difference associated with a twofold change in odds is $B \log 2$. Setting this difference equal to 20 gives

$$B \log 2 = 20, \qquad B = \frac{20}{\log 2}.$$

The analyst could choose a different number of points. The important point is that the score scale becomes interpretable: a fixed number of score points corresponds to a fixed multiplicative change in default odds.

The constant A is chosen by selecting a reference probability and a reference score. Let p_0 be the reference probability and S_0 the score assigned to that probability. Then A is defined by

$$A = S_0 + B \log\left(\frac{p_0}{1 - p_0}\right). \tag{9.3}$$

This makes Equation (9.2) produce Score $= S_0$ when $p_i = p_0$.

Example: constructing a score and assigning a band.
Suppose the organization wants a 20-point increase in score to halve the odds of default. Then

$$B = \frac{20}{\log 2} \approx 28.85.$$

Suppose it also wants an account with default probability 20% to have a score of 650. Thus $p_0 = 0.20$ and $S_0 = 650$. Equation (9.3) gives

$$A = 650 + 28.85 \log\left(\frac{0.20}{0.80}\right) \approx 610.0.$$

Now consider an account with predicted default probability $p_i = 0.30$. Equation (9.2) gives

$$\text{Score}_i = 610.0 - 28.85 \log\left(\frac{0.30}{0.70}\right) \approx 634.$$

If the organization defines the bands as low risk below 10%, moderate risk from 10% to 20%, high risk from 20% to 40%, and very high risk above 40%, then $p_i = 0.30$ places this account in the high-risk band. The score of 634 is not a separate risk estimate; it is the points-scale representation of the same predicted probability.

The probability, the score, and the band have different purposes. The probability p_i is the statistical estimate. The score Score_i translates that probability into an operational scale. The band converts the continuous estimate into a managerial category. The relationship is:

$$p_i \quad \longrightarrow \quad \text{Score}_i \quad \longrightarrow \quad \text{Band}_i.$$

Risk bands are useful because many organizations do not want a different policy for every possible probability. They want a small number of action categories. For example:

$$\text{Band}_i = \begin{cases} \text{Low risk}, & p_i < 0.05, \\ \text{Moderate risk}, & 0.05 \le p_i < 0.15, \\ \text{High risk}, & 0.15 \le p_i < 0.30, \\ \text{Very high risk}, & p_i \ge 0.30. \end{cases}$$

These cutoffs are illustrative. In practice, they should be chosen using calibration evidence, exposure, severity, review capacity, risk appetite, pricing policy,

and governance requirements. They may be defined directly in probability space or translated into equivalent score intervals using Equation (9.2).

A band is therefore not merely a label. It is a decision category. A low-risk band may support standard acceptance and monitoring. A high-risk band may trigger manual review. A very-high-risk band may trigger escalation, restructuring, exposure reduction, or rejection. The purpose of the score is to make these decisions more consistent, transparent, and monitorable.

Table 9.1: From model output to risk decision

Model output	**Operational transformation**	**Risk-management use**
Predicted probability	Probability score or scorecard points	Ranking, pricing, provisioning, review
Expected loss	Exposure-sensitive score	Portfolio prioritization and resource allocation
Feature effects	Reason codes or decision narrative	Human review, challenge, documentation
Risk bands	Policy categories	Approval, monitoring, escalation, intervention

Scores should also be embedded in risk assessment. The model estimates likelihood, but the organization must still define the scenario, assess the consequence, judge the quality of the evidence, and decide whether the residual risk is acceptable [7]. This is why calibrated scoring and managerial judgment must be combined. The model helps quantify risk; the risk-assessment process decides what to do with that quantification.

9.3 Calibration and Probability Reliability

Section 9.2 showed how probabilities can be transformed into scores and bands. This section asks whether those probabilities are reliable enough to support that transformation. Calibration matters because a score band is only as credible as the probability estimates behind it. If the model systematically overstates or understates event probabilities, the resulting bands, thresholds, expected-loss estimates, and pricing rules may all be misleading.

Calibration means that predicted probabilities correspond to observed frequencies. If many comparable accounts receive probabilities close to 10%, then approximately 10% of those accounts should experience the event. Calibration is therefore essential when probabilities are used for pricing, provisioning, expected-loss analysis, capital allocation, or risk appetite monitoring.

Let p_i be the predicted probability assigned to case i, and let Y_i be the event indicator. Calibration can be expressed as

$$\Pr(Y_i = 1 \mid p_i = q) = q. \tag{9.4}$$

Here q is a generic probability level, such as 0.10 or 0.30. The equation means that among cases assigned probability q, the event frequency should be q. This is a population statement, not a statement about a single case. A single borrower either defaults or does not default. Calibration can only be assessed over groups of comparable predictions [2, 3].

In practice, exact groups with identical probabilities are rare. Predictions are therefore grouped into bins. A bin is a group of cases with similar predicted probabilities. Let B_k denote the set of cases in bin k. The mean predicted probability in bin k is

$$\bar{p}_k = \frac{1}{|B_k|} \sum_{i \in B_k} p_i, \tag{9.5}$$

and the observed event rate is

$$\bar{y}_k = \frac{1}{|B_k|} \sum_{i \in B_k} Y_i. \tag{9.6}$$

A reliability diagram plots $\bar{y}_k$ against $\bar{p}_k$. A well-calibrated model lies close to the 45-degree line.

The expected calibration error summarizes the bin-level gaps:

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{n} \left| \bar{y}_k - \bar{p}_k \right|. \tag{9.7}$$

Here K is the number of bins and n is the number of observations in the validation or test sample used to compute the statistic. The term $|\bar{y}_k - \bar{p}_k|$ is the calibration gap in bin k, and $|B_k|/n$ weights that gap by the size of the bin. ECE is useful, but it should not replace reliability diagrams or band-level validation. A model can have a small average calibration error while still being poorly calibrated in the high-risk region where decisions are made.

Calibration is different from ranking. A model can rank borrowers well while assigning probabilities that are systematically too high or too low. Ranking may be sufficient for a review queue. Calibration is needed when probabilities enter monetary or regulatory calculations.

9.3.1 Post-hoc Calibration: Logistic and Isotonic Maps

Post-hoc calibration methods adjust model outputs after the predictive model has been trained. They are useful when a model ranks cases well but produces probabilities that are too high, too low, or unevenly calibrated. The base model remains unchanged. A calibration map is learned on validation data and then applied to future predictions.

It is important to distinguish three quantities. The raw model output is denoted z_i. It may be a logit, a margin, or an uncalibrated probability produced by the model. It is not the final operational score Score_i. The original model probability is p_i. The calibrated probability is p_i^{cal}. The purpose of calibration is to replace p_i by p_i^{cal} when the latter has better probability reliability.

For example, a gradient-boosting model may produce an internal raw margin z_i. The software then transforms that margin into a probability p_i. If the resulting probabilities are too confident, a calibration method can learn a new mapping from z_i or p_i to a calibrated probability. This is a technical probability adjustment. It is different from the managerial transformation that converts p_i into Score_i and then into a risk band.

Logistic calibration, often called Platt scaling, fits a logistic mapping from the raw model output z_i to the event indicator Y_i:

$$p_i^{\text{cal}} = \frac{1}{1 + \exp[-(a + bz_i)]}. \tag{9.8}$$

The parameters a and b are estimated on validation data. Logistic calibration is useful when the miscalibration pattern is approximately S-shaped. It is simple, stable, and often preserves ranking because the mapping is monotone when $b > 0$ [8].

Isotonic regression is more flexible. It estimates a non-decreasing function $m(\cdot)$ such that

$$p_i^{\text{cal}} = m(z_i), \qquad m(z_i) \leq m(z_j) \text{ whenever } z_i \leq z_j. \tag{9.9}$$

The word "isotonic" means order-preserving. If one case has a higher raw model score than another, isotonic calibration does not allow it to receive a lower calibrated probability. The fitted function is often stepwise. This flexibility can be useful when the calibration curve is not well described by a logistic shape, but it can overfit when the calibration sample is small [13, 11].

Example: an isotonic calibration map.
Suppose validation data show the following observed default rates by raw-score range:

Raw score range	Observed default rate
0.00–0.20	0.12
0.20–0.40	0.28
0.40–0.70	0.55

An isotonic calibration map may assign

$$m(z) = \begin{cases} 0.12, & 0.00 \le z < 0.20, \\ 0.28, & 0.20 \le z < 0.40, \\ 0.55, & 0.40 \le z \le 0.70. \end{cases}$$

A future case with raw score $z_i = 0.35$ would therefore receive calibrated probability $p_i^{\text{cal}} = m(0.35) = 0.28$. The mapping is monotone: higher raw scores do not receive lower calibrated probabilities.

The practical workflow is straightforward. Train the base model on the training set. Generate raw scores or probabilities on the validation set. Fit the calibration map on the validation set. Apply the map to the test set. Keep the calibrated probabilities only if they improve probability reliability on held-out data. Calibration is not an automatic ritual; it is an empirical adjustment that must earn its place.

Calibration is especially important after resampling or class weighting. Oversampling, undersampling, and class weighting can improve rare-event detection, but they may distort the relationship between predicted probabilities and real deployment frequencies. For this reason, calibration should be checked on data that preserve the deployment base rate.

9.4 Thresholds, Expected Loss, and Portfolio Consequences

A score becomes operational when it is connected to a policy. The policy may be a threshold, a review queue, a pricing rule, a monitoring band, or an escalation rule. The same probability can lead to different actions depending on exposure, severity, capacity, and risk appetite.

Probability ranking is useful when all cases have similar consequences. In many risk settings, however, consequences differ across cases. Missing a

default on a large exposure is not the same as missing a default on a small exposure. Let a_i denote exposure and g_i denote the loss rate or severity conditional on the event. The expected loss for case i is

$$e_i = p_i g_i a_i. \tag{9.10}$$

This is the same expected-loss logic used in Chapter 8. It changes the ranking problem. A high-probability, low-exposure case may generate less expected loss than a moderate-probability, high-exposure case.

The expected-loss calculation also helps evaluate review capacity. Let Q_α denote the top α share of cases ranked by probability score or expected loss. The share of total expected loss captured by that group is

$$e\text{-Share}(\alpha) = \frac{\sum_{i \in Q_\alpha} e_i}{\sum_{i=1}^{n} e_i}. \tag{9.11}$$

This diagnostic answers a managerial question: how much predicted loss is concentrated in the cases the organization can actually review?

Thresholds should also reflect asymmetric costs. Let C_{FP} denote the cost of a false positive and C_{FN} the cost of a false negative. In a credit setting, a false positive is a non-defaulting account that is flagged, reviewed, restricted, repriced, or declined unnecessarily. A false negative is a defaulting account that is not flagged and therefore remains unmanaged.

If a case is flagged, the cost occurs only when the case would not have defaulted. The expected cost of flagging is therefore

$$(1 - p_i)C_{\text{FP}}.$$

If the case is not flagged, the cost occurs only when the case defaults. The expected cost of not flagging is

$$p_i C_{\text{FN}}.$$

The case should be flagged when the expected cost of flagging is no larger than the expected cost of not flagging:

$$(1 - p_i)C_{\text{FP}} \leq p_i C_{\text{FN}}.$$

Solving this inequality gives the cost-sensitive threshold:

$$\tau^* = \frac{C_{\text{FP}}}{C_{\text{FP}} + C_{\text{FN}}}. \tag{9.12}$$

A case is flagged when $p_i \geq \tau^*$. This result follows from comparing the conditional expected cost of action with the conditional expected cost of inaction [4]. It also explains why a default threshold of 0.5 is rarely appropriate in risk management. If missing a default is much more costly than reviewing or restricting a non-default, the threshold should be below 0.5.

The term "review cost" should be interpreted broadly. In an AI-supported workflow, the marginal cost of producing an automated recommendation may be low, but the cost of acting on a false alert is not zero. It may include customer friction, unnecessary limit reductions, adverse-action communication, compliance review, exception handling, and the opportunity cost of scarce human attention. For this reason, even AI-assisted credit systems still require explicit threshold and capacity design.

ROC curves and precision–recall curves remain useful because they show how error rates change as the threshold changes [5, 1]. However, the final threshold is not chosen by a curve alone. It reflects risk appetite, review capacity, expected loss, customer friction, and governance constraints.

9.5 Model Selection, Interpretability, Monitoring, and Governance

Model selection in risk scoring is not a contest for the highest AUC. The analyst is selecting a decision system. A model with strong ranking performance may be inadequate if its probabilities are poorly calibrated. A model with strong probability diagnostics may be inappropriate if it cannot be explained in a regulated setting. A model with high recall may fail operationally if it generates more alerts than the organization can act upon responsibly.

The relevant object is the full scoring pipeline: preprocessing, feature construction, model estimation, calibration, thresholding, score-band definition, explanation, monitoring, and governance. A logistic scorecard may be preferred when interpretability and stability dominate. A gradient-boosting model may be preferred when its improvement in ranking and probability quality justifies the additional validation and explanation burden. A simpler model may be preferred when operational reliability or documentation is the binding constraint.

Interpretability is part of this selection problem. Logistic regression and shallow trees are interpretable by design. More complex models may require post-hoc explanations, variable-importance summaries, partial-dependence plots, or reason codes. These tools are useful, but they do not eliminate model risk. Explanations should be stable, plausible to domain experts, and

reliable in the high-risk region where decisions are made [9, 6, 10].

Monitoring completes the scoring system. After deployment, the organization should track input drift, score drift, event-rate drift, calibration drift, expected-loss drift, override rates, review outcomes, and changes in exposure or severity assumptions. A model may remain statistically stable while the economic meaning of the score changes because exposures grow, recoveries deteriorate, or capacity constraints tighten.

Table 9.2: Governance evidence for a deployed risk score

Governance area	**Evidence to document**
Purpose	Intended use, decision context, users, and prohibited uses
Data	Sources, time period, exclusions, missing-data treatment, target definition
Validation	Train–validation–test design, temporal validation, benchmark models
Calibration	Reliability diagram, ECE, band-level event rates, probability diagnostics
Thresholds	Cost assumptions, capacity constraints, risk appetite, approval authority
Interpretability	Reason codes, feature effects, explanation stability, expert review
Monitoring	Drift metrics, reporting frequency, trigger limits, escalation process
Limitations	Known weaknesses, populations not covered, assumptions, override rules

The model-selection question is therefore managerial as much as statistical: which pipeline produces scores that are accurate enough, calibrated enough, explainable enough, and operationally reliable enough for the decision it supports?

9.6 Empirical Case Study: From Default Prediction to Credit Risk Score

The empirical case study is the applied bridge between Chapters 7, 8, and 9. Chapter 7 used the Taiwan credit-card default dataset to explain classification trees as interpretable segmentation tools. Chapter 8 used the same setting to compare logistic regression, random forests, and gradient boosting as

probability prediction models. This chapter turns the selected probability model into a score-based credit-risk assessment.

The dataset contains 30,000 credit-card accounts and a binary indicator of default in the following month [12]. As before, $Y_i = 1$ means that account i defaults and $Y_i = 0$ otherwise. The predictors include credit limit, demographic variables, repayment-status variables, bill amounts, payment amounts, and constructed summary variables. The same train–validation–test logic is used: training estimates the models, validation supports tuning and calibration checks, and the test set is reserved for final reporting.

The case study proceeds in seven steps: carry forward the predictive benchmark, audit calibration, transform probabilities into scores, define score bands, evaluate review capacity, make the threshold a risk decision, and attach risk-assessment evidence to the score bands.

Step 1: Carry forward the predictive benchmark

Table 9.3 summarizes the test-set performance used as the starting point. The interpretable tree is retained from Chapter 7 as the transparent benchmark. Logistic regression is retained as the classical scorecard baseline. Random forests and gradient boosting are retained because Chapter 8 showed that ensemble models improved ranking and probability quality in this dataset. The Brier score and log-loss were introduced in Chapter 8 as probability-quality diagnostics; they are reported here only to carry forward the model comparison.

Table 9.3: Predictive benchmark carried forward from Chapters 7 and 8

Model	**Mean p**	**AUC**	**Brier**	**Log-loss**
Interpretable tree, depth 3	0.220	0.743	0.138	0.442
Logistic regression	0.220	0.724	0.145	0.467
Random forest	0.221	0.779	0.135	0.430
Gradient boosting	0.222	0.782	0.134	0.428

The table illustrates why scoring should not begin from a single metric. The tree remains useful because it is easy to communicate. Logistic regression remains useful because it is a disciplined scorecard baseline. Gradient boosting is selected for the scoring extension because, in this run, it provides the strongest combination of ranking and probability-quality diagnostics.

Step 2: Audit calibration before scoring

The first scoring question is whether the predicted probabilities can be used as probabilities. Table 9.4 compares the original gradient-boosting probabilities with logistic and isotonic calibration fitted on the validation set.

Table 9.4: Calibration audit for the selected gradient-boosting model

Probability output	**Mean p**	**AUC**	**Brier**	**ECE**
Original gradient boosting	0.222	0.782	0.134	0.015
Logistic calibrated	0.224	0.782	0.134	0.015
Isotonic calibrated	0.224	0.781	0.135	0.016

The audit is informative because it does not show a dramatic gain from post-hoc calibration. Logistic calibration leaves the ranking unchanged and does not improve the Brier score or ECE. Isotonic calibration slightly weakens the test-set diagnostics in this run. The selected scoring model therefore uses the original gradient-boosting probabilities. Calibration remains essential, but it should be retained as an adjustment only when it improves the probability evidence for the intended use.

Table 9.5 reports selected decile-level calibration diagnostics. Accounts are sorted by p_i and divided into ten equal-sized groups. The table reports selected deciles to show the structure of the score distribution.

Table 9.5: Selected decile-level calibration diagnostics

Risk group	**Accounts**	**p range**	**Mean p**	**Default**	**e share**
Lowest 10%	600	2.6–6.0%	5.0%	4.0%	5.5%
Middle decile 5	600	12.0–14.3%	13.1%	15.7%	6.4%
Middle decile 6	600	14.3–17.1%	15.7%	17.3%	6.2%
High 70–80%	600	22.7–33.8%	27.4%	26.8%	12.6%
High 80–90%	600	33.8–57.5%	42.6%	43.7%	15.4%
Highest 10%	600	57.5–89.3%	71.7%	69.5%	25.8%

The columns in Table 9.5 serve different purposes. The p range shows the interval of predicted probabilities inside each decile. Mean p is the model's average predicted probability in that group. The default column is the observed default rate in the test data. Comparing mean p with the observed default rate gives a calibration check. The e share column adds the economic dimension: it reports the share of total predicted expected loss concentrated in that group.

The pattern is encouraging. The lowest-risk decile has mean predicted probability 5.0% and observed default rate 4.0%. The highest-risk decile has mean predicted probability 71.7% and observed default rate 69.5%. The middle and high deciles are not perfectly aligned, as expected in finite samples, but the ordering and approximate probability levels are credible. The table also shows why default probability is not the whole story: the highest decile accounts for 25.8% of predicted expected loss.

Step 3: Transform probabilities into scores

The selected probability output from Step 2 provides the p_i values. Each p_i is converted into a numerical score using Equation (9.2). The score is the continuous number Score_i, not the band. The band is a category created after the probability and score have been calculated.

The case uses the same score-scale convention introduced in Section 9.2: a 20-point increase in score halves the odds of default. Therefore,

$$B = \frac{20}{\log 2} \approx 28.85.$$

The score scale is anchored by assigning a score of 650 to an account with default probability 20%. Thus $p_0 = 0.20$ and $S_0 = 650$. Equation (9.3) gives

$$A = 650 + 28.85 \log\left(\frac{0.20}{0.80}\right) \approx 610.0.$$

The score transformation used in the case study is therefore

$$\text{Score}_i = 610.0 - 28.85 \log\left(\frac{p_i}{1 - p_i}\right).$$

This transformation does not change the ordering of accounts. It converts probabilities into a points scale in which higher scores mean lower risk and score differences have an odds interpretation.

Step 4: Convert scores into risk bands

The case defines four probability bands: low risk below 10%, moderate risk from 10% to 20%, high risk from 20% to 40%, and very high risk above 40%. These bands could equivalently be expressed as score intervals, because Equation (9.2) maps each probability to a unique score.

Table 9.6 reports the resulting band analysis. The "Mean score" column is the average value of Score_i within each band. The band itself is the

managerial category. The score is the numerical scale used to place accounts into that category.

Table 9.6: Score-band analysis for the selected model

Band	**Accounts**	**Mean p**	**Default**	**Mean score**	**Mean e**
Low	1,807	7.0%	6.4%	675	8,373
Moderate	2,186	14.2%	15.4%	652	9,926
High	1,061	28.4%	26.7%	627	18,076
Very high	946	62.7%	62.1%	584	32,761

Table 9.6 is the core managerial output of the case. Mean p is the model's average predicted default probability in each band. Default is the observed default rate in the test data. Their closeness provides a band-level calibration check. The low-risk band has mean $p = 7.0\%$ and observed default 6.4%; the very-high-risk band has mean $p = 62.7\%$ and observed default 62.1%. This gives the score bands a credible probability interpretation.

The mean score column shows the operational score scale. Higher scores correspond to lower default risk: the low-risk band has a mean score of 675, while the very-high band has a mean score of 584. The mean e column adds the economic dimension. The very-high-risk band has the largest average expected loss per account, 32,761, which is almost four times the low-risk band. This is why the band should not merely be monitored; it may require escalation, restructuring, exposure reduction, or declining new credit. The high band also matters: its default rate is far below the very-high band, but its mean expected loss is already roughly twice the moderate band.

The expected-loss column uses Equation (9.10). Because the dataset does not contain realized loss-given-default, exposure a_i is proxied by the credit limit and $g_i = 50\%$ is used as a transparent scenario assumption. The values should therefore be read as decision-support quantities, not as audited accounting provisions.

Step 5: Evaluate review and intervention capacity

A useful score must support capacity decisions. Capacity should be interpreted broadly. It may refer to human credit officers, AI-assisted investigation workflows, customer-contact capacity, compliance review, or the institution's willingness to create friction for customers. Even when AI reduces the cost of screening, the cost of acting on a false alert remains real.

Table 9.7 reports what happens when accounts are sorted by p_i and the organization can act only on the top 5%, 10%, or 20% of accounts.

Table 9.7: Capacity-constrained review diagnostics

Capacity	Reviewed	Default rate	Defaults captured	e captured
Top 5%	300	78.3%	17.7%	13.9%
Top 10%	600	69.5%	31.4%	25.8%
Top 20%	1,200	56.6%	51.2%	41.1%

This table converts model performance into an operating decision. If the team acts on only 300 accounts, the queue is highly concentrated: 78.3% of reviewed accounts default in the test data. But this captures only 17.7% of all observed defaults and 13.9% of predicted expected loss. Expanding action to 1,200 accounts captures just over half of observed defaults and 41.1% of expected loss, but at the cost of a larger intervention burden and more false alerts. The score therefore does not choose the policy. It makes the trade-off visible.

Step 6: Make the threshold a risk decision

Chapter 7 used classification thresholds to teach performance metrics. In a scoring system, the threshold should reflect asymmetric costs and risk appetite. Using Equation (9.12), Table 9.8 shows how the threshold changes as the cost of missing a default rises relative to the cost of flagging a non-default.

Table 9.8: Cost-sensitive threshold illustration

Cost ratio	τ	Flagged	Prec.	Rec.	FP	FN
$C_{\mathrm{FN}}/C_{\mathrm{FP}} = 1$	0.500	703	67.1%	35.6%	231	855
$C_{\mathrm{FN}}/C_{\mathrm{FP}} = 2$	0.333	1,216	56.0%	51.3%	535	646
$C_{\mathrm{FN}}/C_{\mathrm{FP}} = 5$	0.167	2,489	38.7%	72.5%	1,527	365
$C_{\mathrm{FN}}/C_{\mathrm{FP}} = 10$	0.091	4,467	27.4%	92.2%	3,244	104

Precision is the share of flagged accounts that actually default. Recall is the share of all defaulting accounts that the policy captures. False positives are accounts flagged by the policy that do not default. They may create unnecessary review, customer friction, reduced limits, or lost business. False negatives are accounts not flagged by the policy that do default. They represent unmanaged credit risk.

The table shows the policy trade-off. A threshold of 0.50 produces a small and precise flagged set, but recall is only 35.6% and 855 defaults are missed. If the cost of missing a default is five times the cost of flagging a non-default, the threshold falls to 0.167. Recall rises to 72.5%, and missed

defaults fall to 365, but false positives increase to 1,527. If the false-negative cost is ten times the false-positive cost, the policy becomes highly aggressive: recall exceeds 90%, but 3,244 non-defaulting accounts are flagged.

This is the managerial meaning of a threshold. It is not only a statistical cutoff. It is a statement about how the institution balances credit losses, customer friction, review burden, and residual risk.

Step 7: Attach risk-assessment evidence

The final step is to translate score bands into risk-assessment evidence. Table 9.9 combines scenario interpretation, likelihood evidence, and residual-risk action.

Table 9.9: Risk-assessment interpretation of score bands

Band	**Scenario**	**Likelihood evidence**	**Residual-risk action**
Low	Routine account	Mean $p = 7.0\%$; default 6.4%	Accept; monitor portfolio drift
Moderate	Visible repayment weakness	Mean $p = 14.2\%$; default 15.4%	Accept with price or limit review if exposure is high
High	Material repayment-risk signal	Mean $p = 28.4\%$; default 26.7%	Manual review, limit reduction, or repayment mitigation
Very high	Severe repayment-risk signal	Mean $p = 62.7\%$; default 62.1%	Escalate; restructure, suspend exposure growth, or decline new credit

This table keeps both mean p and default because they answer different questions. Mean p reports what the model predicts. Default reports what was observed in the test data. Presenting both gives the reader a direct band-level calibration check. If the two columns were far apart, the policy recommendations would be less credible. Here, the close alignment strengthens the case for using the bands as risk categories.

The residual-risk actions become progressively stronger across the bands. The low band supports acceptance and monitoring. The moderate band supports pricing or limit review when exposure is high. The high band supports manual review or mitigation. The very-high band supports escalation, restructuring, suspension of exposure growth, or declining new credit. The table therefore turns the statistical score into a risk-assessment tool.

The case completes the progression of Part III. Chapter 7 identified interpretable risk segments. Chapter 8 estimated default probabilities and

expected-loss inputs. Chapter 9 turns those outputs into calibrated, banded, monitored, and action-oriented risk decisions.

9.7 Summary and Managerial Takeaways

This chapter showed how predictions become risk decisions. A probability model is only the beginning. To be useful in management, the output must be calibrated, translated into a score, grouped into meaningful bands, connected to thresholds and capacity, and monitored after deployment.

The chapter's central message is that a risk score is a decision instrument. It must be meaningful, reliable, actionable, and governable. A model can rank cases well and still be a poor risk-scoring system if its probabilities are miscalibrated, its thresholds ignore costs, or its bands do not correspond to feasible actions.

Managerial takeaways

- A probability becomes useful only when it is connected to a decision rule.
- Calibration is the evidence that a score can be trusted as a probability input.
- A score scale translates probabilities into an operational language; bands translate that language into action categories.
- Expected loss changes the ranking of priorities when exposure and severity differ across cases.
- Thresholds express risk appetite: lower thresholds capture more risk but create more false alerts.
- Capacity matters even with AI, because acting on a false alert still has business, customer, and governance costs.
- A deployed score must be monitored as both a model and a policy: probabilities, losses, thresholds, overrides, and outcomes can all drift.

Bibliography

[1] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, 2006.

[2] A. Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

[3] Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32(1/2):12–22, 1983.

[4] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pages 973–978, 2001.

[5] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[6] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[7] Nicola Paltrinieri, Louise Comfort, and Genserik Reniers. Learning about risk: Machine learning for risk assessment. *Safety Science*, 118:475–486, 2019.

[8] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA, 1999.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings*

of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144, 2016.

[10] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

[11] Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112:3211–3260, 2023.

[12] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

[13] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.

