Business Analytics

Classification Evaluation:
ROC/PR, Thresholding,
Cost-Sensitive, Calibration

Why This Matters

- Decisions depend on evaluation, not only accuracy
- Asymmetric costs: fraud, churn, health triage
- Imbalance & capacity constraints demand careful metrics

Agenda

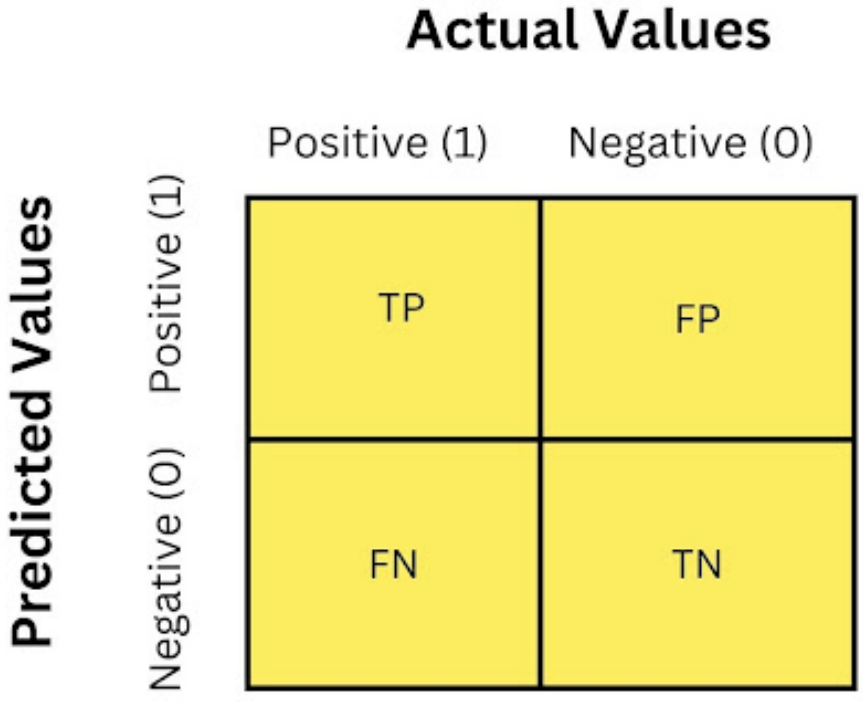
- Confusion matrix & metrics refresher
- ROC & AUC
- PR & Average Precision
- Threshold selection
- Cost-sensitive evaluation
- Probability calibration
- Case study + exercises

Learning Objectives

- Explain ROC & PR and when to use each
- Choose thresholds for different objectives
- Incorporate unequal costs into decisions
- Assess and improve calibration

Confusion Matrix Refresher

- TP, FP, TN, FN definitions
- Predicted vs Actual 2×2 layout
- Foundation for derived metrics



Core Rates

- TPR/Recall/Sensitivity = TP/(TP+FN)
- FPR = FP/(FP+TN); TNR/Specificity
- Precision (PPV) and NPV

F-measure Family

- General formula for "F-measure Family" is called the F-beta score (or F_β)
- Calculates the harmonic mean of Precision and Recall, and it allows to adjust the weight between them using the beta β parameter.

$$\text{Recall} = \frac{TP}{TP+FN}$$

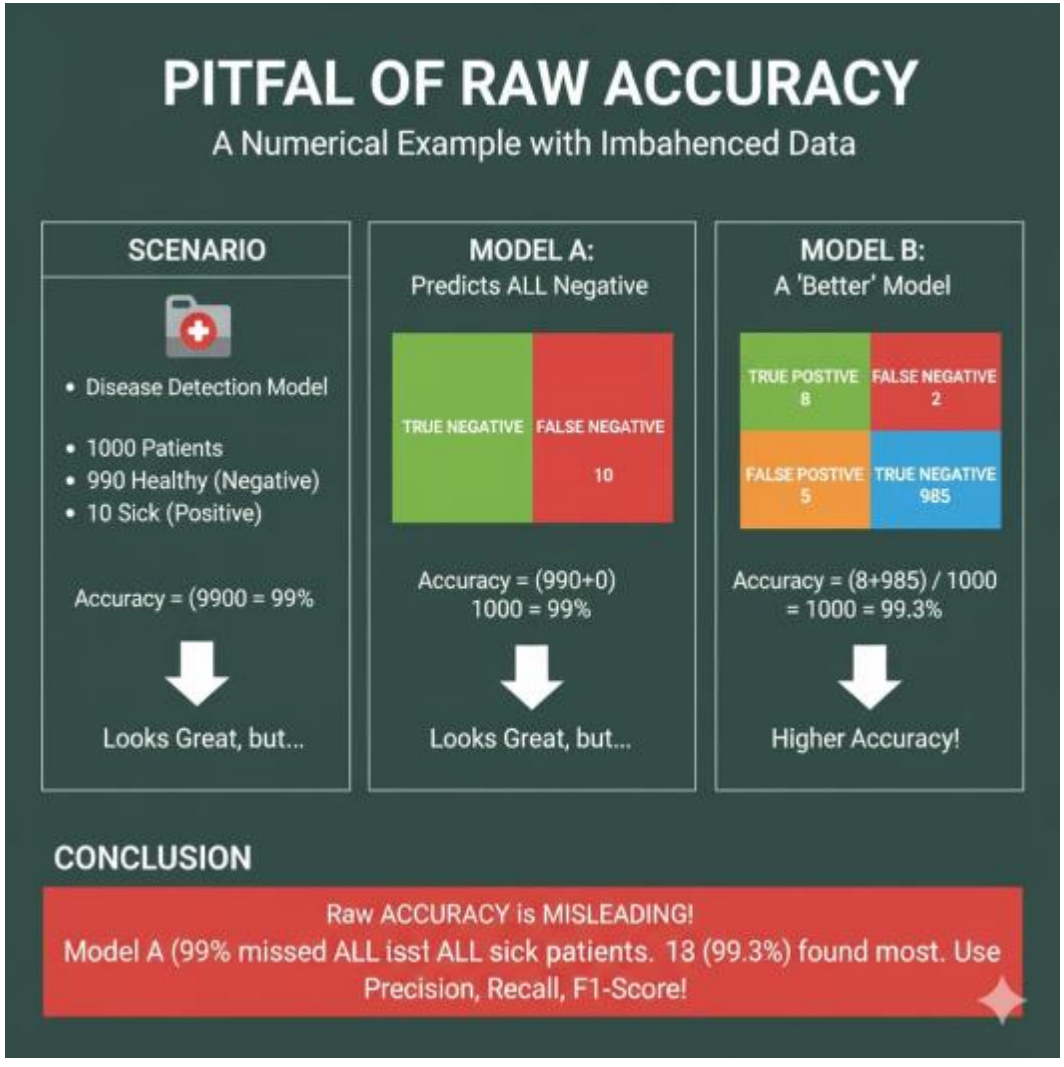
- If β = 1 (F1-Score): Precision and Recall are considered equally important.

$$\text{F1} = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$$

- Useful for early model selection

Accuracy vs Class Imbalance

- High accuracy can be misleading when positives are rare
- Always predicting negative may look good but is useless
- Prefer class-sensitive metrics and curves



Matthews Correlation Coefficient (MCC)

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- Robust single-number metric even under imbalance
- Range [−1, 1]; 0 ~ random, 1 ~ perfect
- **Example (Imbalanced Case):** 1,000 samples, of which 990 are *Negative* and 10 are *Positive*. A "lazy model" simply predicts all 1,000 samples as *Negative*.
  - **TN = 990, FP = 0**
  - **FN = 10, TP = 0**
  - **Accuracy:** (990 + 0) / 1000 = **99%** (looks very high!)
  - **F1-Score: 0** (because TP = 0).
  - **MCC: 0** (or undefined, depending on how the denominator is handled), indicating the model has no predictive power at all.

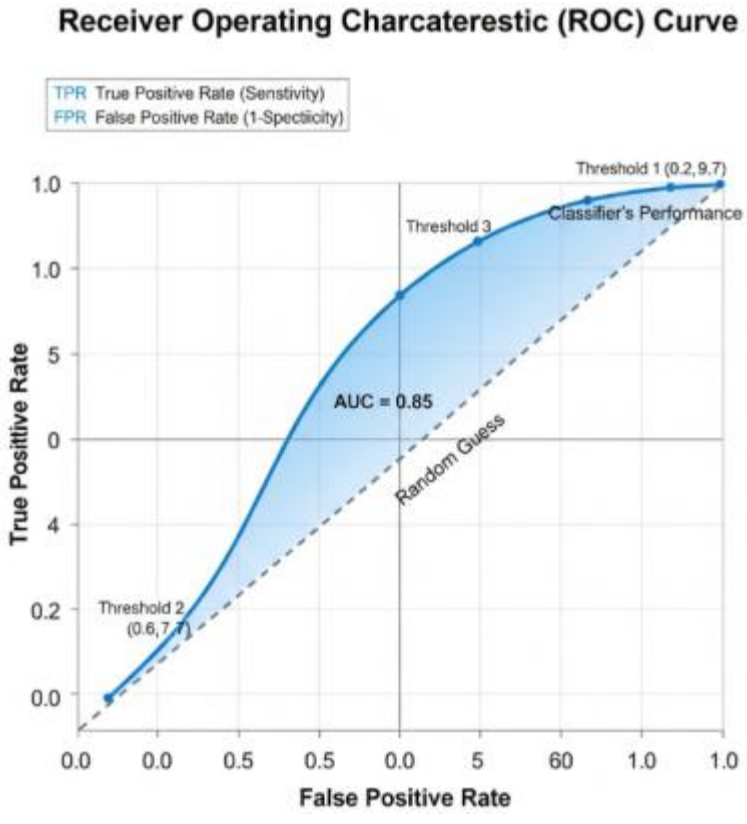
Balanced Accuracy & G-Mean

- Balanced Accuracy calculates the arithmetic mean of Sensitivity and Specificity

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

- Balanced Accuracy = (TP/(TP+FN) + TN/(TN+F))/2
- If a "lazy model" predicts everything as the majority class (e.g., the Negative class), its Sensitivity = 0 and Specificity = 1 ➔ Balanced Accuracy (0 + 1)/2 = 0.5 (or 50%), indicating performance is no better than random guessing (unlike the 99% raw accuracy).
- G-Mean calculates the geometric mean of Sensitivity and Specificity.
- G-Mean = sqrt(TPR·TNR)
- G-Mean measures the balance between the performance on the two classes. Avoid majority-class dominance.

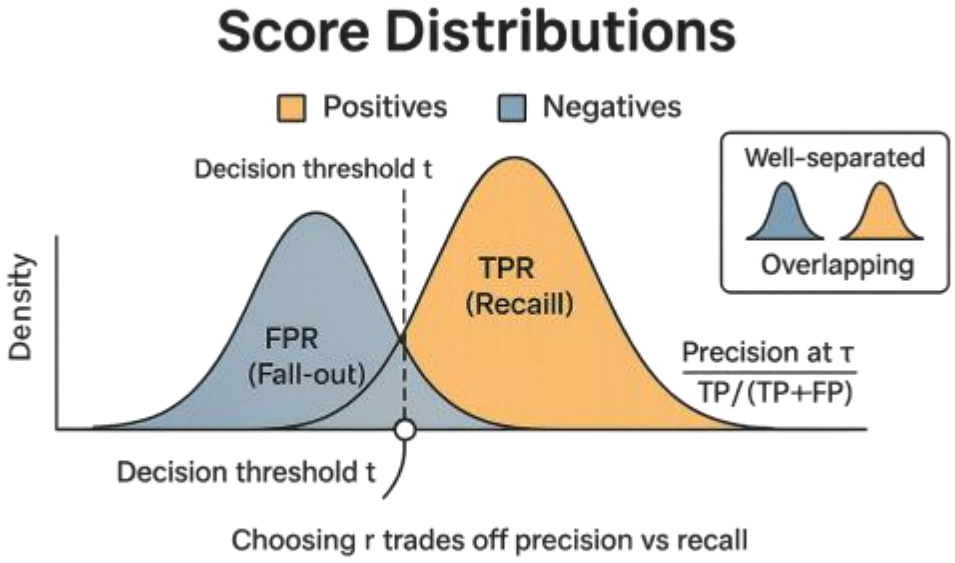
ROC — Concept



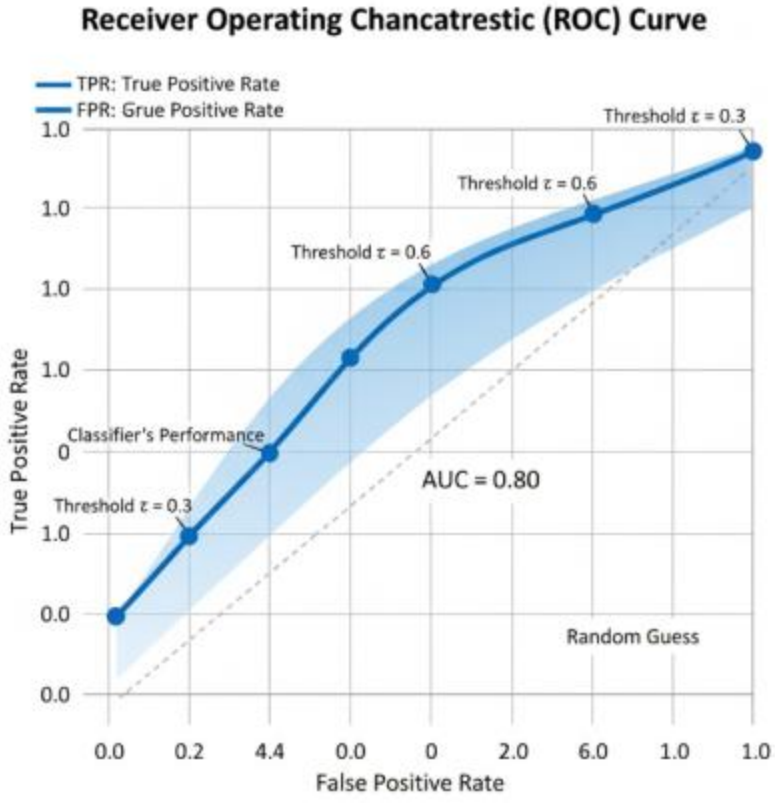
- **ROC: Receiver Operating Characteristic**
- ROC curve is created by taking the **Score Distributions** and trying **every possible threshold** from 0.0 to 1.0
- Plot TPR vs FPR across thresholds
- Above-diagonal indicates better than random
- AUC-ROC: ranking ability

Score Distributions

- Overlapping score histograms for positives/negatives
- Sweeping a threshold traces ROC/PR points
- A confusion matrix is **computed *after* picking a decision threshold** for the model's scores/probabilities.
- Change the threshold ⇒ change which cases are labeled Positive vs. Negative ⇒ the four counts (TP, FP, FN, TN) change—so the confusion matrix "moves."



ROC — Example Table



- Show TP/FP/TN/FN at $\tau \in \{0.8, 0.6, 0.4, 0.3\}$
- Derive TPR/FPR from counts

AUC Interpretation

- AUC (Area Under the Curve) = P(score+ > score−)
- AUC is the probability that a randomly selected positive sample will be ranked higher than a randomly selected negative sample

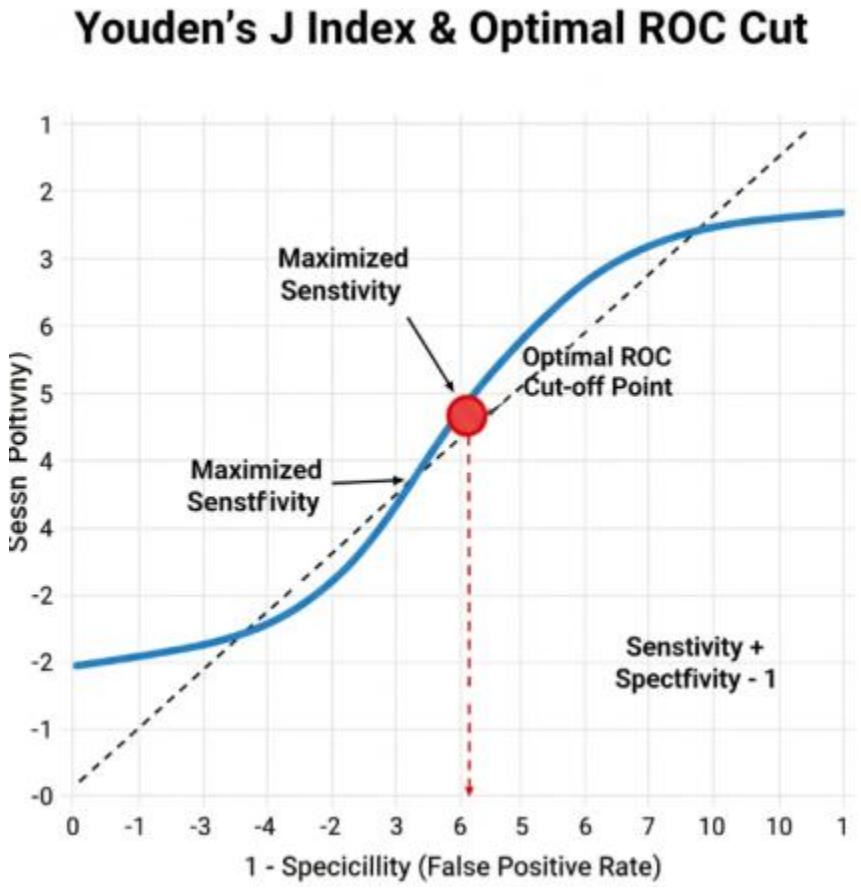
$$AUC = \frac{\sum points}{Total\ Positives\ * Total\ Negatives}$$

  - If Score(Positive) > Score(Negative): Add 1 point
  - If Score(Positive) = Score(Negative): Add 0.5 points
  - If Score(Positive) < Score(Negative): Add 0 points
- Invariant to monotonic transforms
- Ignores prevalence and costs

Youden’s J & ROC Cut

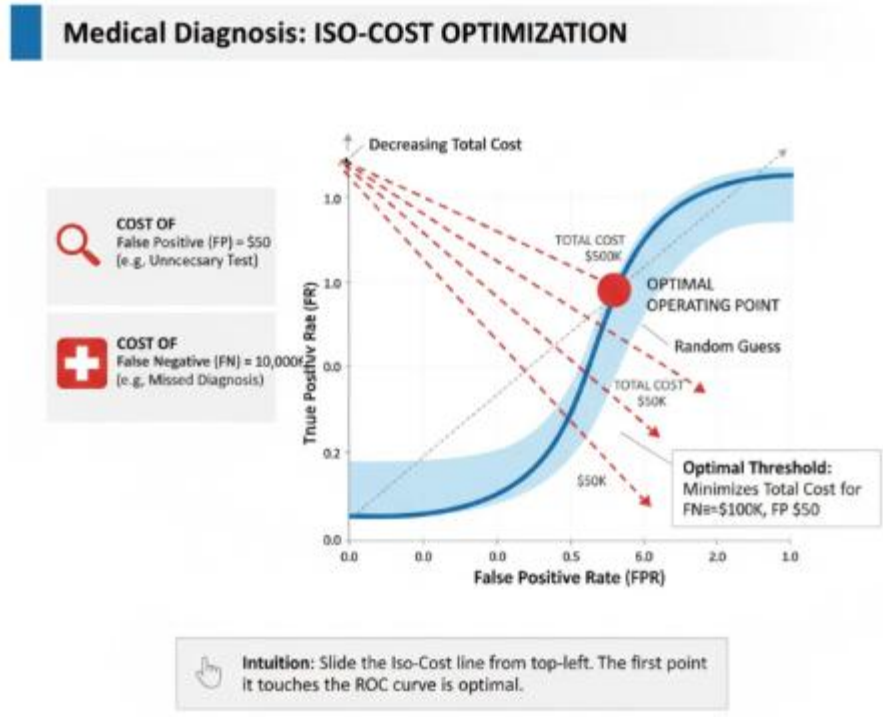
- Youden's J Index is a single metric used to summarize the performance of a classification model.

  J = Sensitivity + Specificity - 1 or J = TPR − FPR
- The value of J ranges from 0 to 1
  - J = 1: a perfect model with no incorrect predictions
  - J = 0: The model has no discriminative ability; its performance is equivalent to a random guess.
  - Maximize when costs equal and prevalence ~50%
- ROC Cut (or "optimal cut-off") refers to the use of the ROC curve to find the best decision threshold for the trained model.
- Youden's J index at any given point on the ROC curve is the vertical distance from that point to the 45-degree diagonal line.
- Optimal cut-off according to Youden's J is the point on the ROC curve with the largest vertical distance from the diagonal line.



Iso-TPR / Iso-Cost Intuition

- An "Iso-TPR" (Equal True Positive Rate) line is just a horizontal line on the ROC plot
- An Iso-Cost line represents all the different ways to achieve the same "Total Mistake Bill"
- **Healthcare Example (Cancer):**
  - **False Positive (FP):** Classifying a healthy person as possibly having the disease.
    Cost: $50 (anxiety/panic cost, plus unnecessary additional tests).
  - **False Negative (FN):** Classifying a sick person as healthy.
    Cost: $10,000 (very high cost due to missed detection and delayed treatment).
- **Spam Filtering Example:**
  - **False Positive (FP):** Sending an important email (e.g., an interview invitation) to the spam folder.
    Cost: $1,000 (very high—missed opportunity).
  - **False Negative (FN):** Letting a spam email into the inbox.
    Cost: $1 (small cost—mostly annoyance)



When ROC Misleads

- Severe imbalance hides FP burden in FPR
- Two ROC-similar points can differ economically
- ROC is often the wrong tool for highly imbalanced classification problems

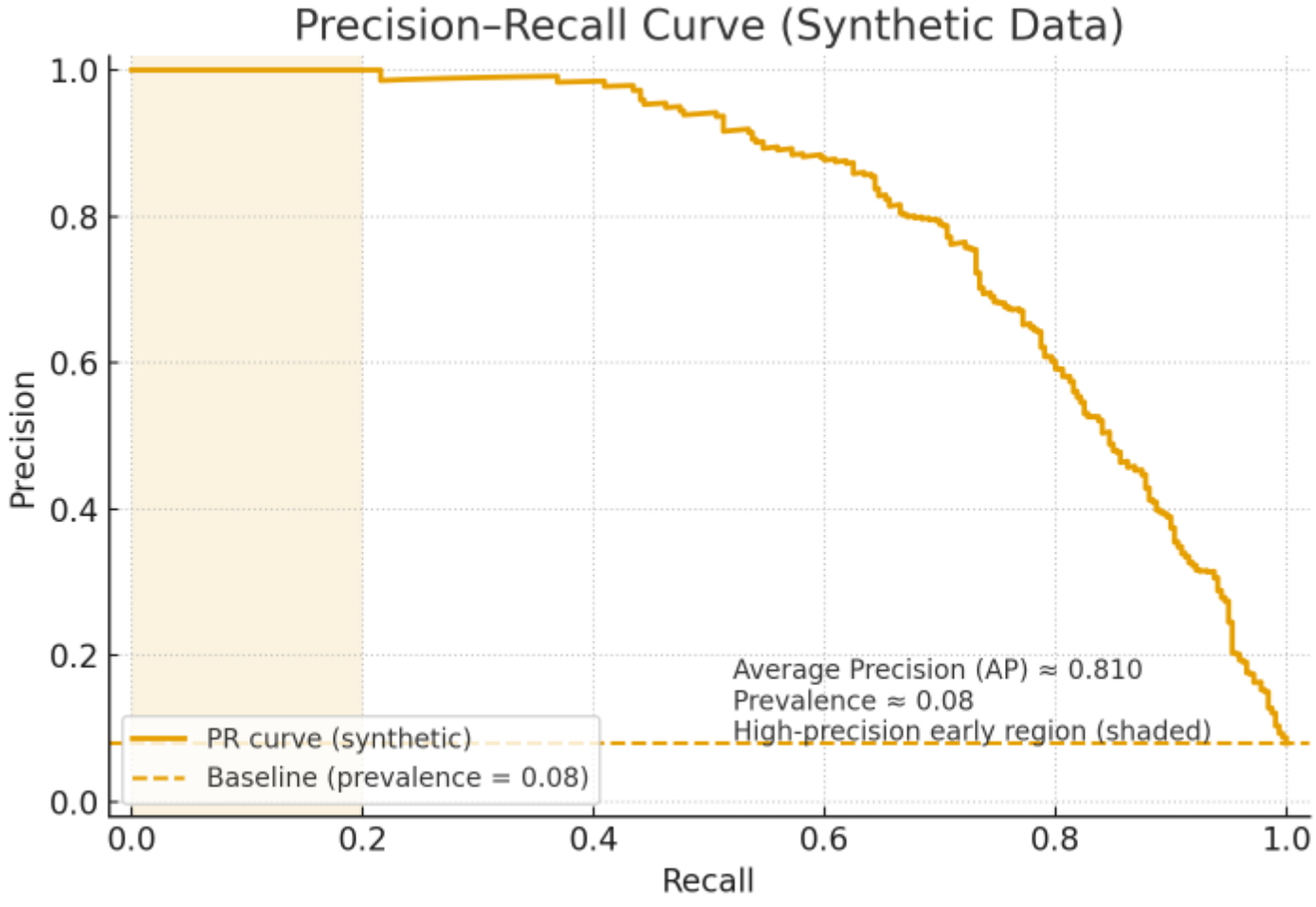
Scenario	Why ROC is Misleading	What to Use Instead
Highly Imbalanced Data	FPR is insensitive; it hides the impact of False Positives, leading to an "overly optimistic" AUC.	**Precision-Recall (PR) Curve**
Unequal Misclassification Costs	ROC treats all errors equally.	**Cost-Benefit Analysis** or manually weighting errors.
Need for Accurate Probabilities	AUC only measures ranking, not probability accuracy.	**Calibration Plots** (e.g., Reliability Diagrams)

PR — Concept

- Precision-Recall (PR) Plot is a model performance evaluation tool for binary classification, useful for imbalanced data.
- PR plot visualizes the trade-off between Precision and Recall at every possible decision threshold ➔ Plot Precision vs Recall across thresholds
- PR curve plots all of these (Recall, Precision) points as you vary the threshold from 0 to 1.

PR Curve (Simulated)

- AP ≈ 0.810 on synthetic data
- Baseline precision ≈ prevalence = 0.08
- Early high-precision region is valuable when review capacity is limited



PR — Example Table

- Compute Precision/Recall at $\tau \in \{0.8, 0.6, 0.4, 0.3\}$
- Baseline precision equals prevalence

threshold (τ)	TP	FP	FN	TN	Precision	Recall
0.8	17	0	23	160	1	0.425
0.6	34	3	6	157	0.9189	0.85
0.4	40	26	0	134	0.6061	1
0.3	40	55	0	105	0.4211	1

Average Precision (AP) vs PR-AUC

- AP is the weighted mean of precision at recall increases
- AP and PR-AUC (Area Under the Precision-Recall Curve) are often used interchangeably and refer to the same concept in classification model evaluation.

$$AP = \sum_{k} (P(k) * \Delta R(k))$$

where P(k) is Precision and $\Delta R(k)$ is the change in Recall at threshold k

- Emphasizes early retrieval quality
- Average Precision (AP) is a precise way to calculate the PR-AUC.

ROC vs PR — When to Use

- ROC: ranking ability; prevalence-invariant
- PR: rare positives; FP costs matter
- Report both for completeness

Threshold by F1/Fβ

- Maximize F1 on validation
  - F1: If your goal is the best possible *balance* between Precision (not flagging non-spam) and Recall (catching all spam), you would tune τ to find the value that gives the highest F1-score.
- For high-risk misses, choose β>1
  - I'd rather flag 10 healthy patients for a follow-up test (False Positives) than miss one person who actually has the disease (a False Negative)."
- Avoid leakage (use held-out data)

Thresholding — Business Framing

- Scores → decisions via a threshold τ
  - Model Score = 0.75 (75% chance of spam)
  - If your threshold $\tau = 0.5$, the decision is "Spam" (because $0.75 > 0.5$).
  - If your threshold $\tau = 0.8$, the decision is "Not Spam" (because $0.75 < 0.8$).
- τ should reflect objectives: F1, profit, recall@k, SLA, regulations

Threshold by ROC Geometry

- Using the shape of this curve to pick the best threshold.
  - Method 1: Closest to the "Perfect" Point (0, 1)
    - Minimizing the Euclidean distance $\sqrt{(FPR-0)^2+(TPR-1)^2}$.This finds a threshold that is "good" at both metrics simultaneously.
  - Equal costs & balanced classes: maximize Youden's J
    - Finding the threshold that gives the highest J = (TPR – FPR) value is considered an "optimal" balanced choice.
- Else use iso-cost slope vs ROC for optimal τ

Threshold by PR Targets

- Constrain precision ≥ P_o and maximize recall
  - Airport security screening or critical medical diagnosis. The goal is to *never miss* a threat or a disease (high Recall), even if it means you have many false alarms (low Precision) that need to be checked manually.
- Or recall ≥ R_o and maximize precision
  - A marketing campaign for a very expensive product. You only want to contact leads you are *at least 90% sure* are interested. You accept you will *miss* some interested leads (low Recall) to avoid annoying uninterested ones (high Precision).

Calibrated Probabilities for Thresholding

- In short: Calibration ensures that a model's 70% confidence score *actually means* there is a 70% chance of the event happening.
  - If you take 100 items that the model gave a score of 0.8 (80%)...
  - ...a perfectly calibrated model means that, in reality, about 80 of those items will be positive.
- Bayes-optimal τ requires well-calibrated probabilities
- Fix calibration first if scores are unreliable

Example: Fraud Risk

- Business Rule: "We need to manually review any transaction with a > 10% chance of being fraud."
- Action: You must set your threshold $\tau = 0.1$.
- If the model is uncalibrated, its 0.1 score might actually mean a 30% real-world risk ➔ Setting threshold wrong and missing the business target ➔ With a calibrated model, can trust that setting τ = 0.1$ correctly implements the 10% risk rule.

Top-k / Quota-Based Thresholding

- The main idea is: "I don't know what the threshold score τ should be, but I know exactly how many items I want to select.“
- Top-k refers to selecting the "k" items with the highest scores.
  - Marketing Budget: "We have a $10,000 budget for a mail campaign, and each mailer costs $2. We can send k = 5,000 mailers."
  - Action: You find the 5,000 customers *most likely* to respond and send mailers only to them.
- Rank by score and take top k or top q%
- Use precision@k, recall@k, lift charts

Cost-Sensitive Evaluation — Setup

- Define the business value or cost for every possible outcome.
- Define cost matrix with benefits for TP

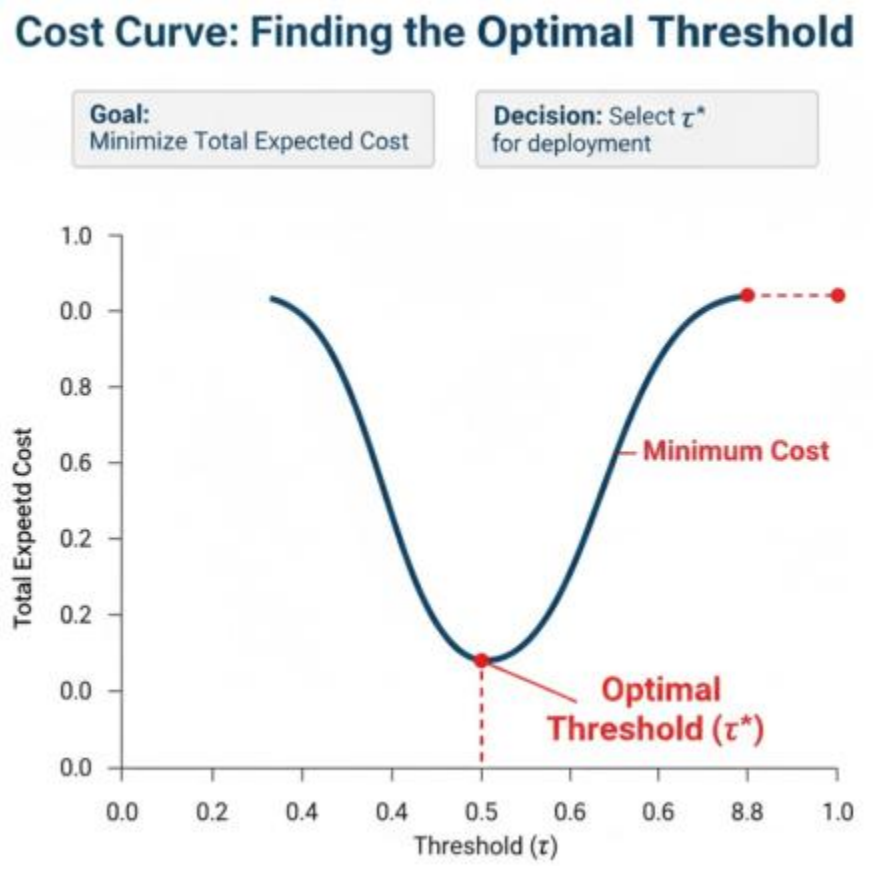
	Predicted: Positive	**Predicted: Negative**
Actual: Positive	**True Positive (TP)**	**False Negative (FN)**
	Value: Often a gain or profit (e.g., +$100 from a sale)	**Cost:** The cost of a missed opportunity (e.g., -$100 in lost profit)
Actual: Negative	**False Positive (FP)**	**True Negative (TN)**
	Cost: The cost of a wasted action (e.g., -$5 for sending a useless ad)	**Value:** Usually $0 (the correct "do nothing" outcome)

Expected Cost at a Threshold

- Expected Cost at a Threshold is used to find the optimal decision point (threshold) for a classification model that minimizes the total cost in a real-world operation.
- It is a combination of two concepts:
  - **Expected Cost:** The average cost you expect to incur for each of the model's predictions.
  - **Threshold:** The "cut-off" point that uses to turn a probability prediction (e.g., 70%) into a concrete decision (e.g., "Yes" or "No").
- Changing the Threshold directly changes the Confusion Matrix, and therefore changes the Total Expected Cost.
- **"Expected Cost at a Threshold"** is the process where you *test* every possible threshold (e.g., 0.01, 0.02, ..., 0.99), and for each threshold:
- Calculate the confusion matrix (TP, FP, TN, FN) at that threshold.
- Calculate the "Total Expected Cost" based on that confusion matrix and your cost matrix.
- When plotting this ➔ curve. **The lowest point (minimum) on this curve is the "Optimal Threshold."**

Cost Curves

- Cost Curve is a graph that visualizes the Total Expected Cost (or Total Expected Profit) across *every* possible threshold τ value



Threshold from Costs (Bayes Rule)

- Optimal threshold τ is the point at which the expected cost of predicting "Positive" (Class 1) is equal to the expected cost of predicting "Negative" (Class 0).
- Assume the cost for correct predictions (True Positive, True Negative) is 0, then the formula for the optimal threshold is:

$$\tau = C_{FP}/(C_{FN} + C_{FP})$$

where**:**

  - τ(Threshold): The optimal threshold.
  - C_{FP} (Cost of False Positive): The cost you incur when you mistakenly predict "Positive" (when it was actually "Negative").
  - C_{FN} (Cost of False Negative): The cost you incur when you mistakenly predict "Negative" (when it was actually "Positive").
- **Decision Rule:**
  - Assuming your model outputs a probability $p = P(Y=1|X)$ (the probability of being "Positive" given the data X):
    - If $p > \tau$ ➔ predict "Positive" (Class 1).
    - If $p < \tau$ ➔ predict "Negative" (Class 0).

Risk & Regulatory Constraints

- Risk" and "Regulatory Constraints" are what determine your costs C_{FP} and C_{FN} and limit how choosing threshold.
- Add constraints: max FPR, min precision, fairness bounds
- Risk of a False Positive C_{FP}
  - *Example (Spam Filter):* The risk is angering a customer who misses an important email. C_{FP} is high.
- Risk of a False Negative C_{FN}
  - *Example (Fraud Detection):* The risk is losing the entire amount of the transaction. C_{FN} is very high.
- Regulatory Constraints ➔ Setting Hard Limits
- Solve as constrained optimization on validation
  - Find the minimum cost point *that satisfies all regulatory constraints*.

Imbalanced Data Considerations

- Report PR, precision@k, calibration
- Stratified CV; keep test distribution
- Resampling/weights for training only

Lift & Gain Charts

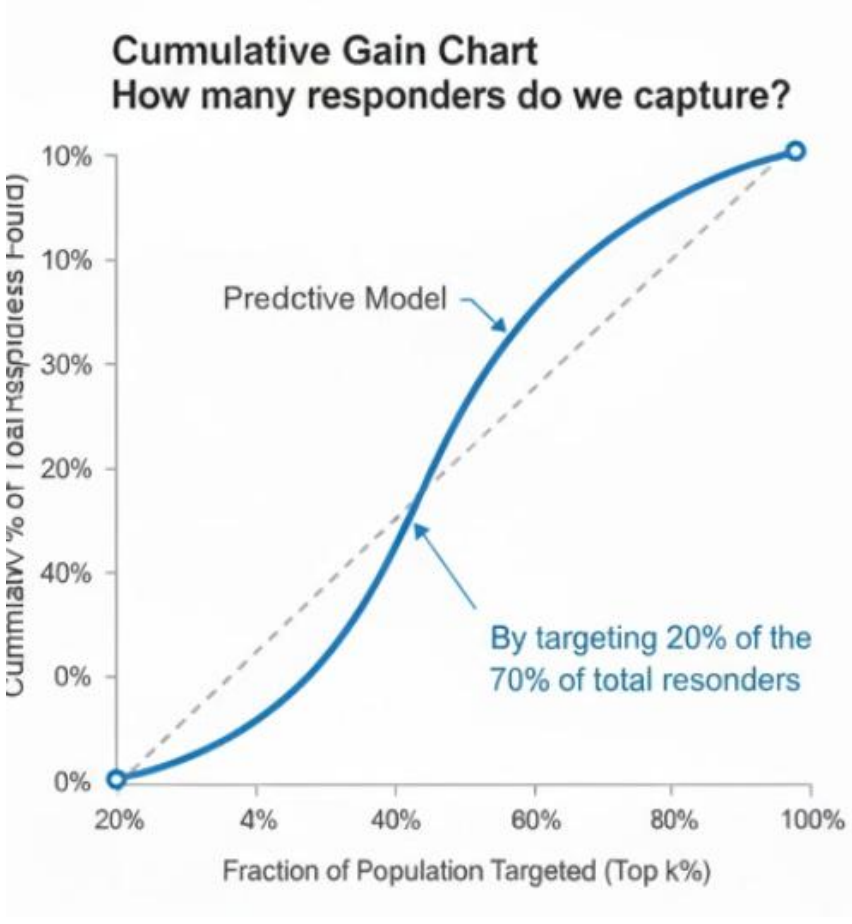
- Cumulative gain and lift vs fraction targeted
  - Gain Chart shows the "cumulative benefit" when targeting a certain percentage of the population, sorted by the model's score (from highest to lowest).
  - Lift Chart directly answers the question: "How many times better is our model than selecting randomly?
- Useful for marketing top-k selection

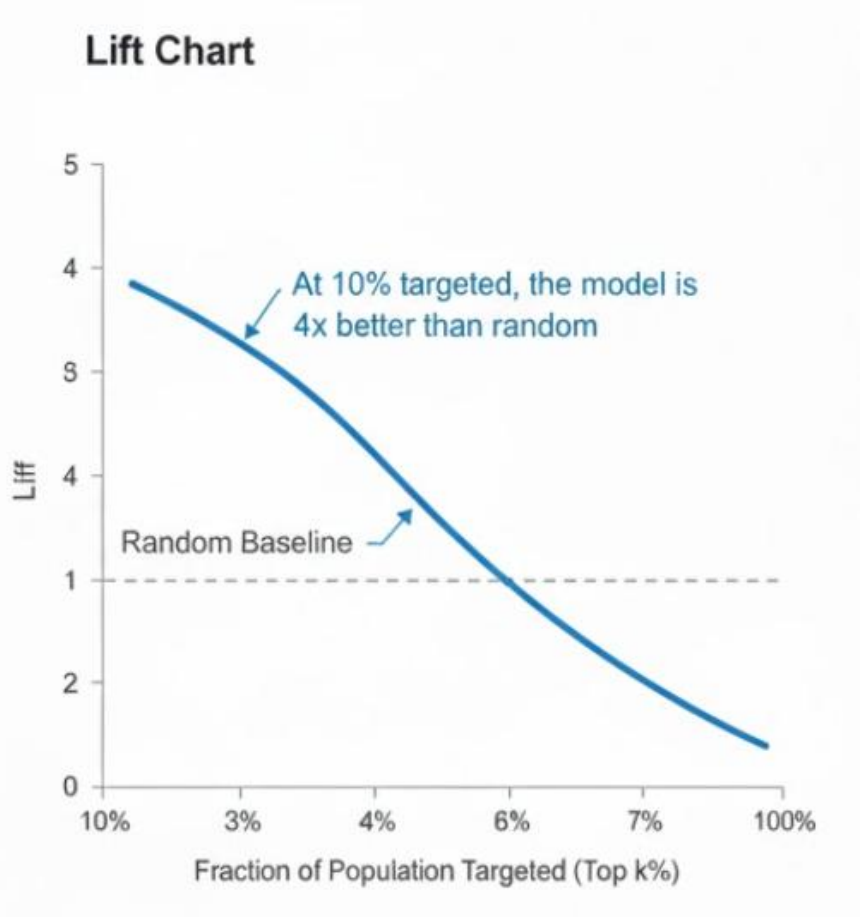
Feature	Cumulative Gain Chart	Lift Chart
Key Question	**How many** responders do we capture?	**How many times better** is our model?
Y-axis	Cumulative % of Positives (e.g., 0% - 100%)	A factor (e.g., 1x, 2x, 5x)
Baseline	45-degree diagonal line	Horizontal line at Y = 1
Useful for	Understanding scale: "Targeting 20% captures 70% of all responders."	Understanding efficiency: "Targeting 20% is 3.5 times better than random."

Cost-Sensitive Learning vs Post-hoc

- In-training: class weights, focal loss, custom utility
- Post-hoc: tune τ using costs
- Often combine both

Lift & Gain Charts





Probability Calibration — Why Care?

- Action rates depend on probability accuracy
- A model that only ranks (like one measured by AUC or a Lift chart) tells that Customer A is more likely to churn than Customer B. It's great for prioritizing (e.g., "call our top 1,000 riskiest customers").
  - But a calibrated model tells you that Customer A has an 80% probability of churning and Customer B has a 30% probability. This allows for much smarter actions:
    - Action for A (80% risk): This is an emergency. Offer a large, expensive discount to save them. The high probability justifies the high cost.
    - Action for B (30% risk): This is a moderate risk. Send a low-cost "we miss you" email. Offering a large discount would be a waste of money—they probably weren't going to leave anyway.

    ➔ If your model is uncalibrated and says both customers have an "80% risk" (a common issue with models like Random Forest, which can be over-confident), you would waste money by giving the expensive discount to Customer B.
- Miscalibration → wrong ROI, wrong capacity

Calibration Diagnostics

- Reliability curve (predicted vs observed)
  - A visualization tool to assess a model's calibration
  - All predicted probabilities (e.g., from 0.0 to 1.0) are divided into a number of "bins" (e.g., 10 bins: [0-0.1], [0.1-0.2], ..., [0.9-1.0]).Within each bin, we calculate two va
- Brier score, Log loss, ECE/MCE
  - Brier score: a "proper scoring rule" used to measure both discrimination and calibration. Essentially, it is the Mean Squared Error (MSE) applied to probability predictions
  - Log loss: also known as **Cross-Entropy Loss**. This is also a "proper scoring rule" and is often used as the loss function when training classification models (like Logistic Regression).
  - ECE/MCE: Both of these metrics are calculated based on the same "bins" used in the Reliability Curve. They summarize the deviation of the reliability curve from the perfect diagonal into a single number.

When to Care (and When Not To)

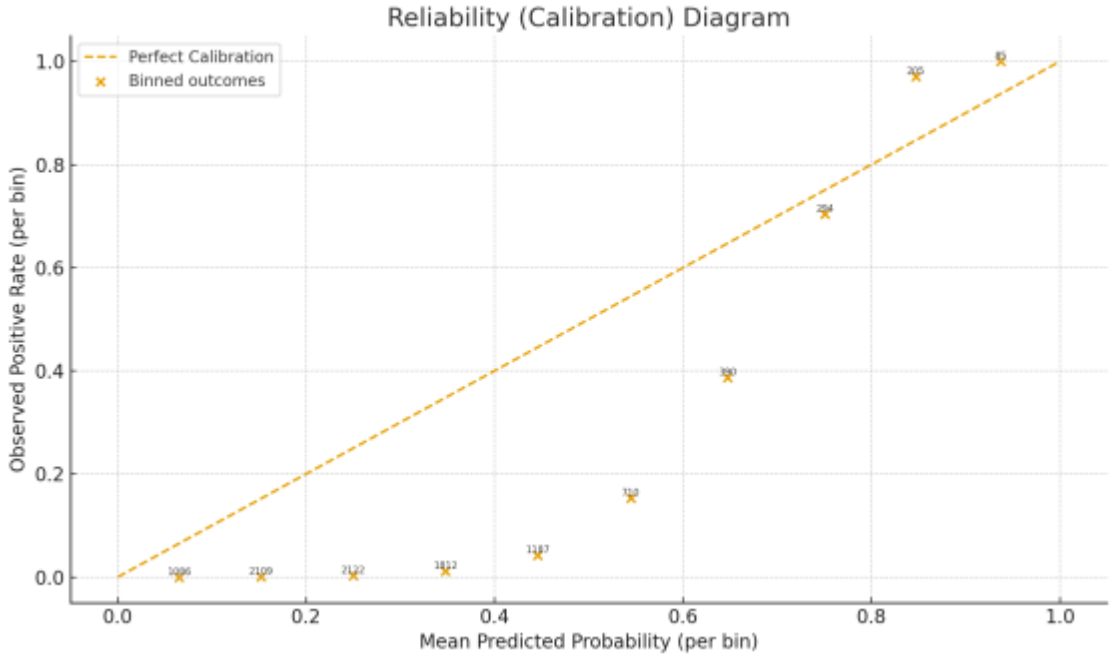
Scenario	Do You Need Calibration?	Why?
Ranking Contest (e.g., Kaggle)	**No**	Only the order of predictions matters (AUC, Lift).
Finding "Top 100" Users	**No**	This is a pure ranking and selection task.
Calculating ROI / Profit	**YES (Critical)**	The actual probability value is a number in your financial formula.
Setting Insurance Premiums	**YES (Critical)**	The premium is a direct function of the predicted risk probability.
Medical Diagnosis	**YES (Critical)**	A doctor needs to know if the risk is actually 80% or 40% to decide on surgery.
Operational Planning	**YES (Critical)**	You base staffing, inventory, and server load (capacity) on how many events you expect.

Calibration Diagnostics

Metric	Meaning	Best is
Reliability Curve	Visualization (Predicted vs. Observed)	On the diagonal $y=x$
Brier Score	MSE for probabilities (Measures calibration & discrimination)	Closer to 0
Log Loss	"Surprise" level (Penalizes confident errors heavily)	Closer to 0
ECE	*Average* calibration error	Closer to 0
MCE	*Worst-case* calibration error	Closer to 0

Reliability Diagram (Simulated)

- 10 bins from 0–1; show observed vs predicted
- Counts per bin provide stability context



Before vs After Calibration

- Show reliability curves and Brier/ECE deltas
  - Before Calibration: The model's curve (e.g., for a Random Forest or Neural Network) often has an "S" shape. This indicates the model is over-confident.
  - After Calibration: The calibration process (e.g., Isotonic Regression) has "learned" this S-curve and "bent" it back
- Demonstrate impact on decision quality
  - Dishonest probabilities lead to bad business decisions.

Metric	Before Calibration (Example)	After Calibration (Example)	Delta (Change)	Meaning
Brier Score	0.135	**0.110**	**-0.025**	The average probability error (MSE) has decreased. The model produces more "accurate" probabilities.
ECE	0.092 (9.2%)	**0.014** (1.4%)	**-0.078**	This is the biggest improvement. The average deviation from the perfect diagonal dropped from 9.2% to just 1.4%.
Log Loss	0.450	**0.390**	**-0.060**	The penalty for confident but wrong predictions has significantly decreased.
AUC	0.850	0.850	**~ 0**	**Important Note:** Calibration (almost) never changes the AUC, as it doesn't change the ranking of the predictions.

Cross-Validation for Calibration

- Hold-out set for calibrator; or out-of-fold predictions
  - Training the calibrator. The calibrator's job is to learn the biases of the main model. To do this, it must see the main model's predictions on data it wasn't trained on
- Evaluate calibration on untouched data
  - This is for evaluating the final, calibrated model. After training the main model (on data A) and training calibrator (on data B), need a completely separate, "untouched" set of data (data C) to see if the entire pipeline works.

Effect of Calibration on Thresholding

- With good calibration, τ^*aligns with profit optimum
  - The Theoretical Path (Calibrated): Use the calculated threshold, τ^*).The Empirical Path (Uncalibrated): Use an empirical cost curve.
- Without it, rely on empirical cost curves

Method	Meaning of Threshold τ∗	How to find τ∗
Calibrated Model	**Theoretical:** The threshold is a real business risk (e.g., 15%).	**Calculation:** Based on costs (e.g., $\tau^* = \frac{cost_{FP}}{cost_{FN}}$)
Uncalibrated Model	**Empirical:** Just an arbitrary number (e.g., 0.28) that minimizes cost on a validation set.	**Search:** Use an empirical cost curve.

Putting It Together — Workflow

- Train → Evaluate (ROC/PR) → Calibrate → Choose τ → Back-test
- Check fairness, capacity, stability

Common Pitfalls

- Reporting AUC only
- Choosing τ on test set (leakage)
- Ignoring capacity constraints
- Not communicating trade-offs

Practical Tips (Ops)

- Fix SRM in experiments before evaluation
- Use stratified splits
- Track drift; recalibrate periodically

Communicating to Stakeholders

- Show ROC & PR plots + operating points
- Cost table and capacity impact for 2–3 τ scenarios
- What-if analysis for prevalence shifts

Evaluation Checklist

- Stratified splits, no leakage
- ROC & PR (AUC/AP)
- Calibration diagnostics (Brier, ECE)
- Validated cost/benefit model
- Threshold with constraints/capacity

Further Reading

- Saito & Rehmsmeier (2015) — PR vs ROC
- Flach (2016) — ROC analysis
- Niculescu-Mizil & Caruana (2005) — Calibration
- scikit-learn User Guide

Implementation Hints (Python)

- sklearn: roc_curve, precision_recall_curve, CalibratedClassifierCV
- Reliability: bin by predicted p and compute observed rates
- Use cross-validation for model & calibrator

Q&A + Thank You

- Questions