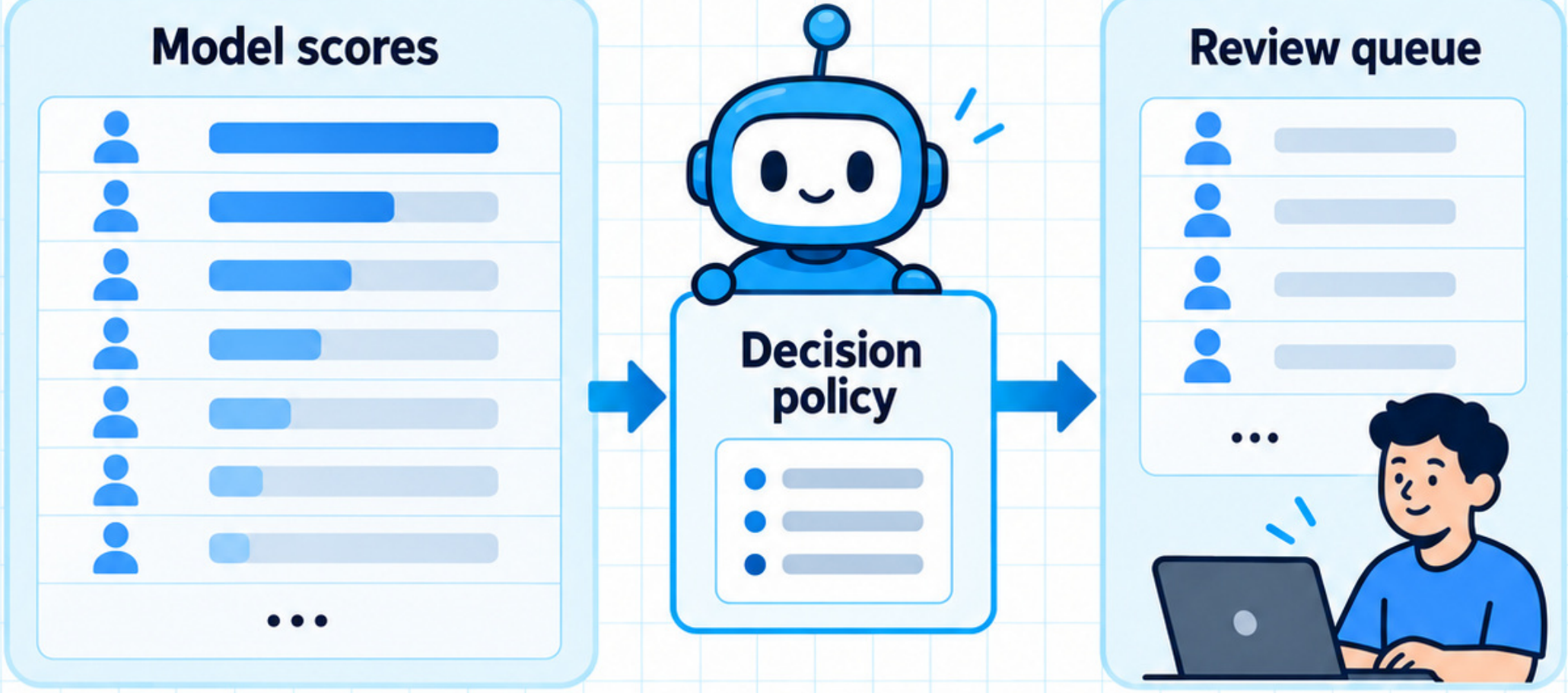
A SCORE STILL NEEDS A DECISION

Baselines and decision thresholds

Our model ranks customers by future purchase inactivity.
A team can review only a few.

Today: train a simple model, compare it with a baseline, and choose a decision rule that fits the job.



FREEZE THE TARGET BEFORE TRAINING

Use the Day 7 target: **label 1** means no completed purchase in the next 30 days, with sufficient follow-up. **Label 0** means at least one.

Available Features

Features must be available at prediction time.

Later Observed Label

Label 1 means no completed purchase in the next 30 days, with sufficient follow-up.

Label 0 means at least one.

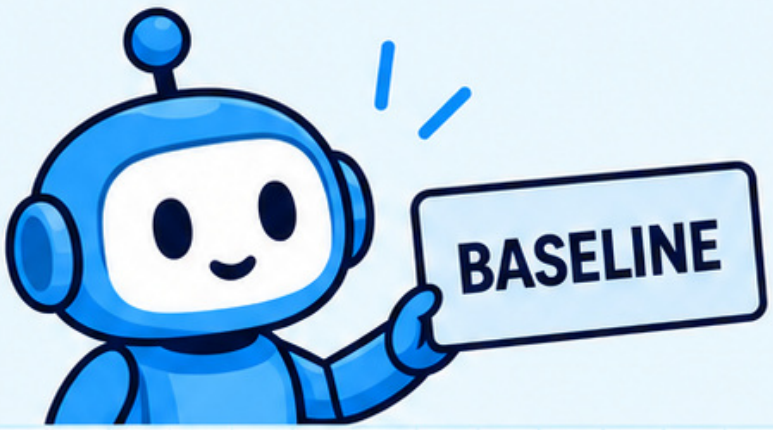
Review Decision

Our decision is review priority.

Predicting inactivity does not prove that contacting someone will help.

GIVE THE LEARNED MODEL A SIMPLE OPPONENT

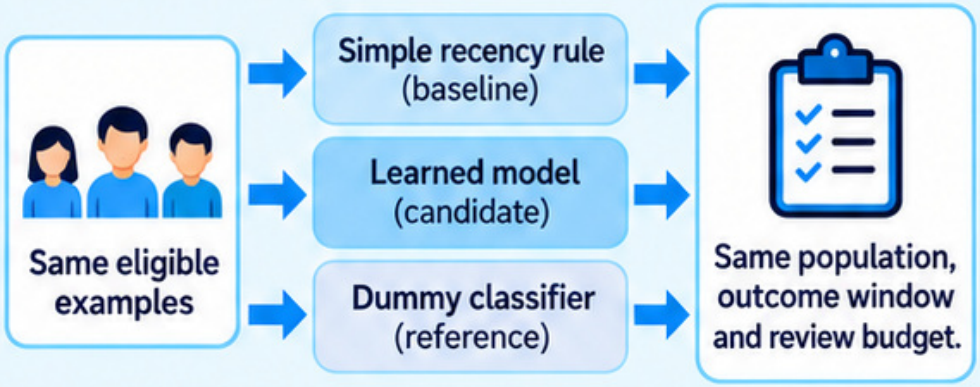
A baseline is a clearly defined reference behavior.



For this queue, a simple rule could rank eligible customers by time since their last completed purchase.

A dummy classifier can expose a weak classification benchmark.

Compare candidates on the same population, outcome window and review budget.



TRAIN ONE COMPLETE PIPELINE

1

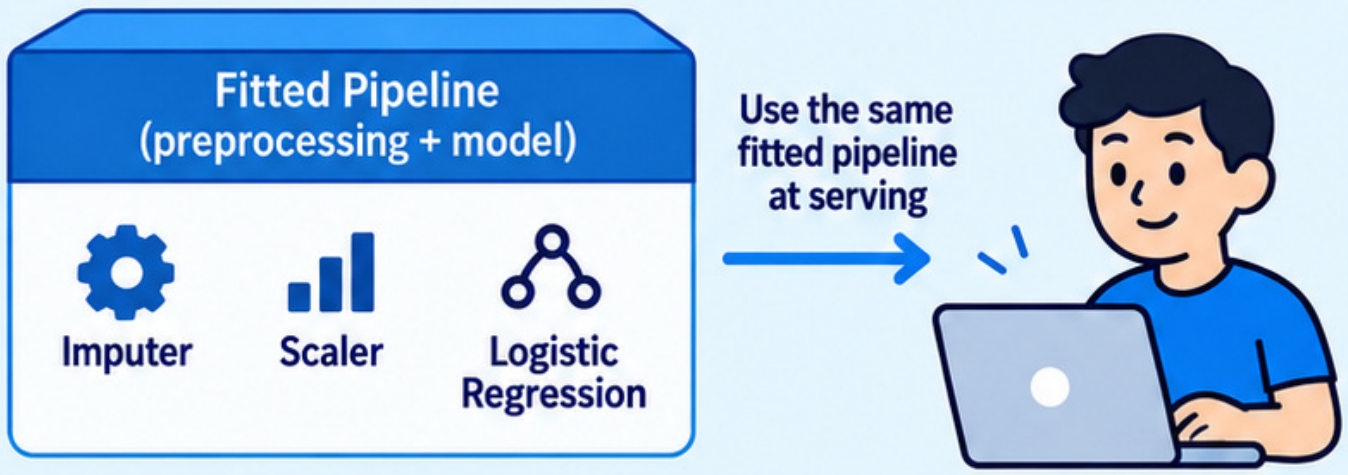
Start with a small feature set and a simple classifier such as logistic regression.

2

Fit missing-value handling, scaling and model parameters using training rows only.

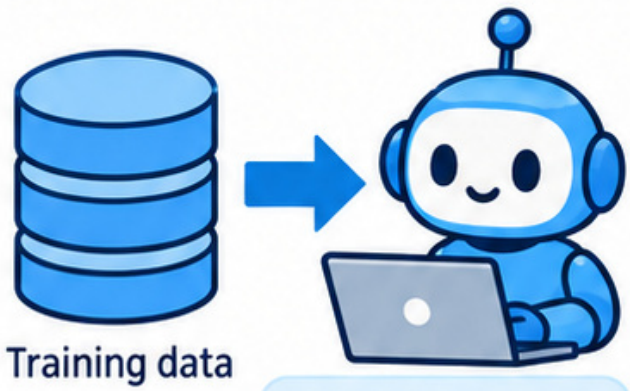
3

Save the fitted preprocessing with the model. The serving path must reuse those learned rules.



SEPARATE LEARNING FROM CHOOSING

Training data fits the pipeline.



Validation data compares models and decision thresholds.

The final test evaluates the frozen choice.

For time-dependent data, preserve the ordered cutoffs and label-availability rules from Day 9. Tuning on the training set or final test weakens the evidence.

INSPECT A SMALL VALIDATION COHORT

Toy validation data: eight customers, four observed positives.
Higher scores mean greater estimated inactivity.

Customer	Score	Observed label (collected after complete follow-up)
A	0.95	1
B	0.88	0
C	0.82	1
D	0.70	1
E	0.65	0
F	0.50	0
G	0.40	1
H	0.20	0

These numbers illustrate decisions; they are not a trained model benchmark.

MAKE THE CUTOFF RULE EXPLICIT

Our toy rule selects a customer when score ≥ threshold.

score ≥ threshold

Define what happens exactly at the threshold.

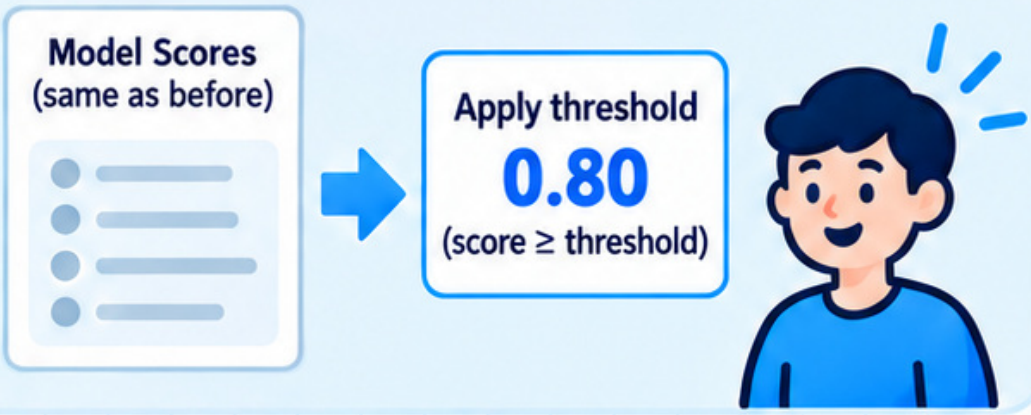
At threshold 0.80, A, B and C enter review. D through H do not.

Customer	Model Score	Selected?
A	0.95	Yes
B	0.88	Yes
C	0.82	Yes
D	0.70	No
E	0.65	No
F	0.50	No
G	0.40	No
H	0.20	No

Threshold 0.80

The model scores have not changed.

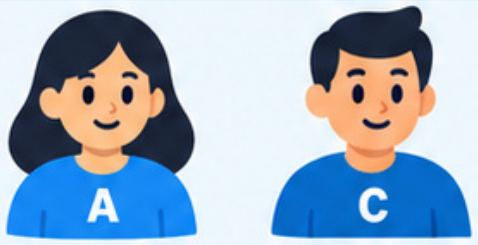
The application changed how scores become selections.



COUNT WHAT THE SELECTION GETS RIGHT

At threshold 0.80 in our toy cohort:

True positives:
A and C → 2



2

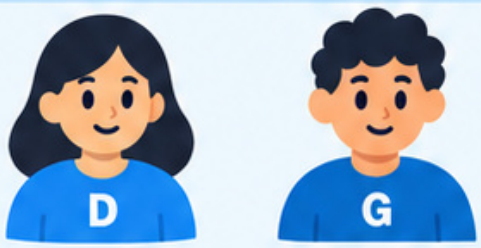
Selected positive
(sent for review, observed inactivity)

False positives:
B → 1

1

Selected negative
(sent for review, not observed inactivity)

False negatives:
D and G → 2



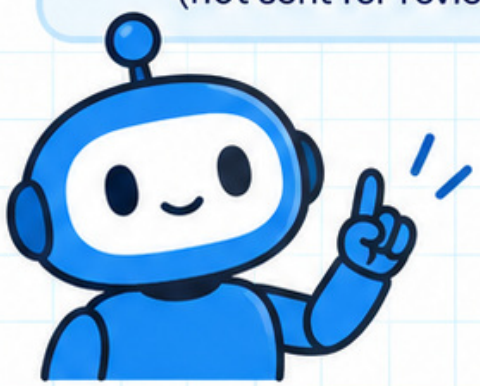
2

Unselected positive
(not sent for review, observed inactivity)

True negatives:
E, F and H → 3

3

Unselected negative
(not sent for review, not observed inactivity)



Positive means observed inactivity.
Selected means sent for review.
Keep those meanings separate.

KEEP THE DENOMINATORS VISIBLE

Precision asks: of the selected customers, how many were positive?

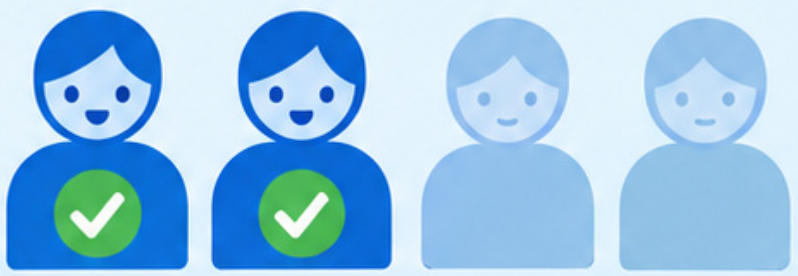
Selected customers
(at threshold 0.80)

3 selected

$$\frac{2}{3} = 66.7\%$$

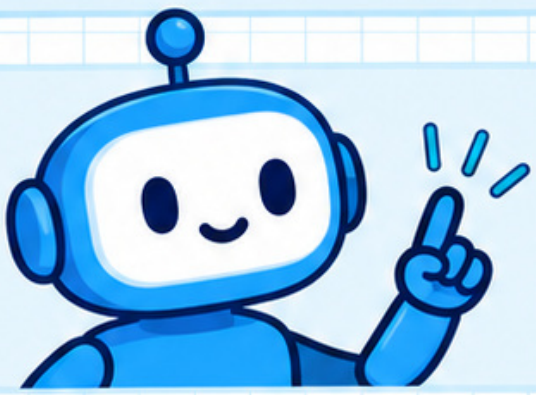
Recall asks: of all observed positives, how many did we select?

All observed positives
(in validation set)



4 actual positives

$$\frac{2}{4} = 50\%$$



These are toy validation results at threshold 0.80. Report the cohort size and the selection rule with the numbers.

A LOWER THRESHOLD CHANGES THE WORKLOAD

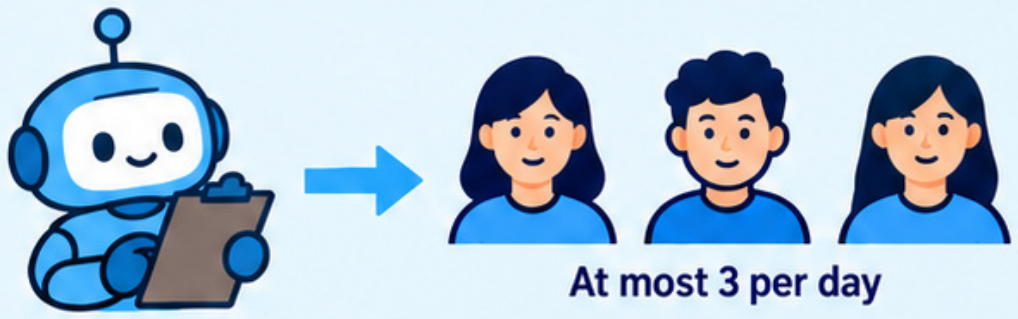
Toy comparison on the same eight customers:

Threshold		Precision	Recall
≥0.80 (toy)	3 selected (toy)	66.7% (toy)	50% (toy)
≥0.60 (toy)	5 selected (toy)	60% (toy)	75% (toy)
≥0.40 (toy)	7 selected (toy)	57.1% (toy)	100% (toy)

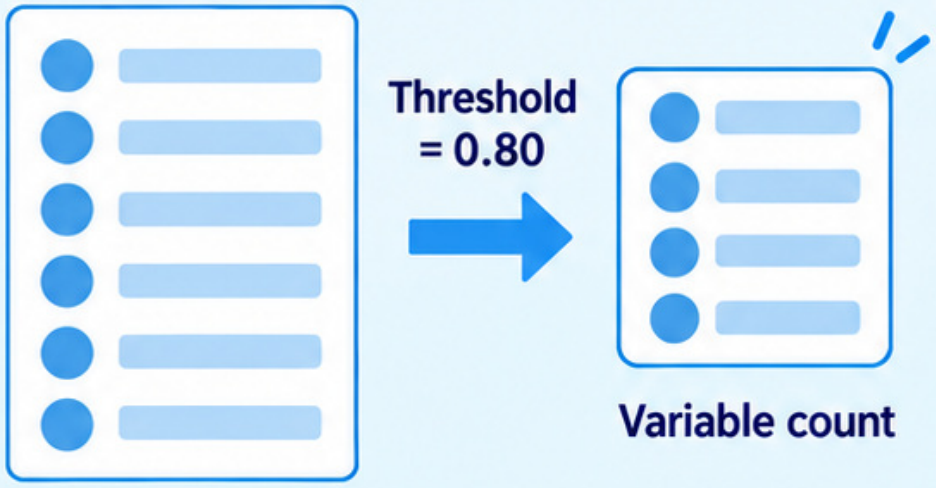
A lower threshold finds more positives here, but also sends more people to review.

A THRESHOLD IS NOT A CAPACITY GUARANTEE

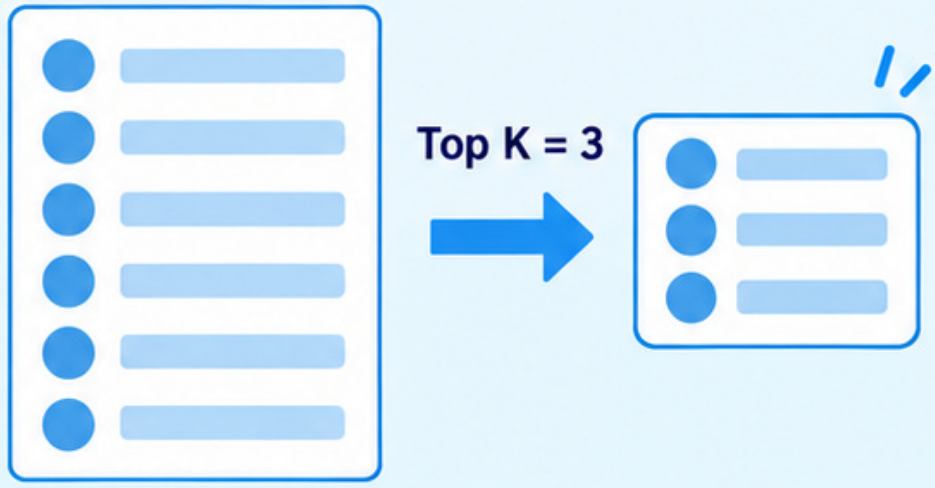
Suppose the team can review at most three customers per day.



Threshold Policy (score cutoff)

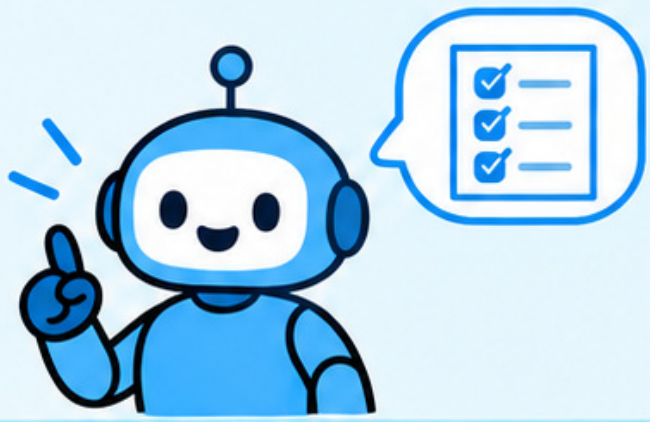


Top-K Policy (caps selected count)



Threshold 0.80 fits this toy cohort.
A different day may produce more than three qualifying scores.

A top-K policy caps the queue; a threshold sets a score cutoff. Define their combination, tie-breaking and overflow behavior deliberately.



COMPARE AT THE SAME REVIEW BUDGET

A, B, C, D, E, F, G, H

Same eight customers (A, B, C, D, E, F, G, H)

Toy capacity: three reviews from the same eight customers.

Recency-rule baseline

selects D, B, E: one positive.

Precision 1/3; recall 1/4.

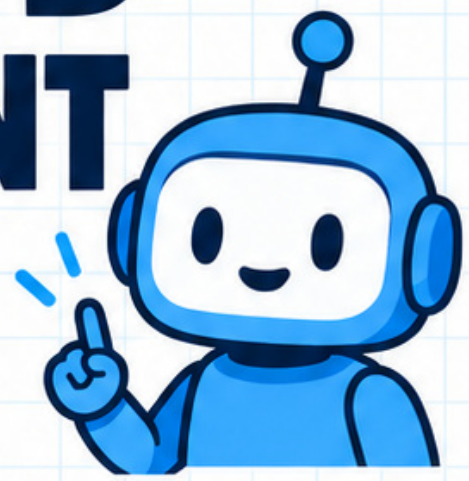
Learned-score ranking

selects A, B, C: two positives.

Precision 2/3; recall 2/4.

This small example explains comparison.
It does not establish production superiority.

ERROR COSTS ADD ANOTHER CONSTRAINT



Toy accounting rule: a false positive costs 1 unit; a false negative costs 3 units.

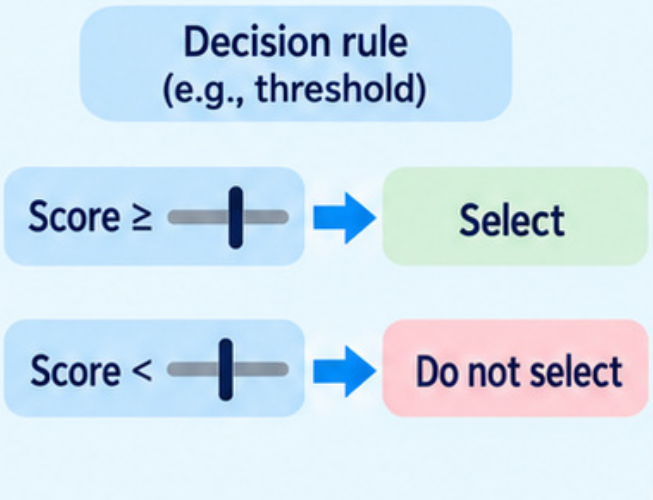
Threshold	False Positives (FP)	False Negatives (FN)	Total Cost (units)	Selected customer count
At threshold **0.80:**	1	2	1 + 3×2 = **7 units.**	✓ Within capacity: 3 reviews ≤ 3
At threshold **0.60:**	2	1	2 + 3×1 = **5 units.**	✗ Exceeds capacity: 5 reviews > 3
At threshold **0.40:**	3	0	3 + 3×0 = **3 units.**	✗ Exceeds capacity: 7 reviews > 3

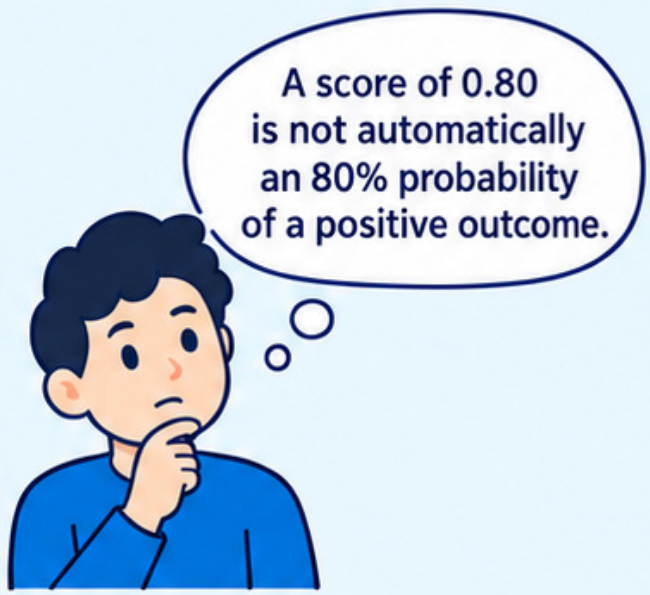
The lower-cost options exceed our three-review budget. Choose under all declared constraints.

A HIGH SCORE IS NOT A GUARANTEE

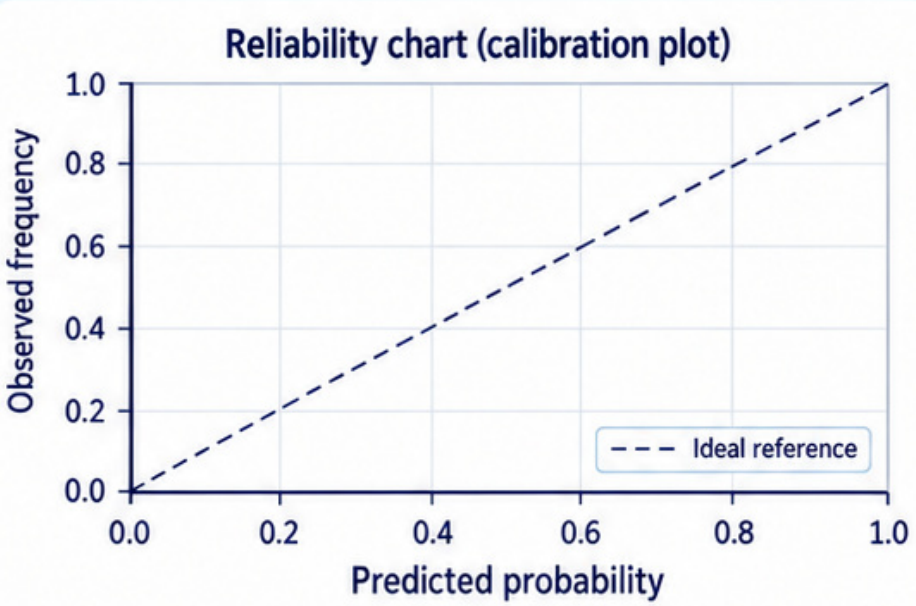
A score supports ranking or a decision rule. Treating 0.80 as an accurate 80% probability needs calibration evidence.

A score helps rank.

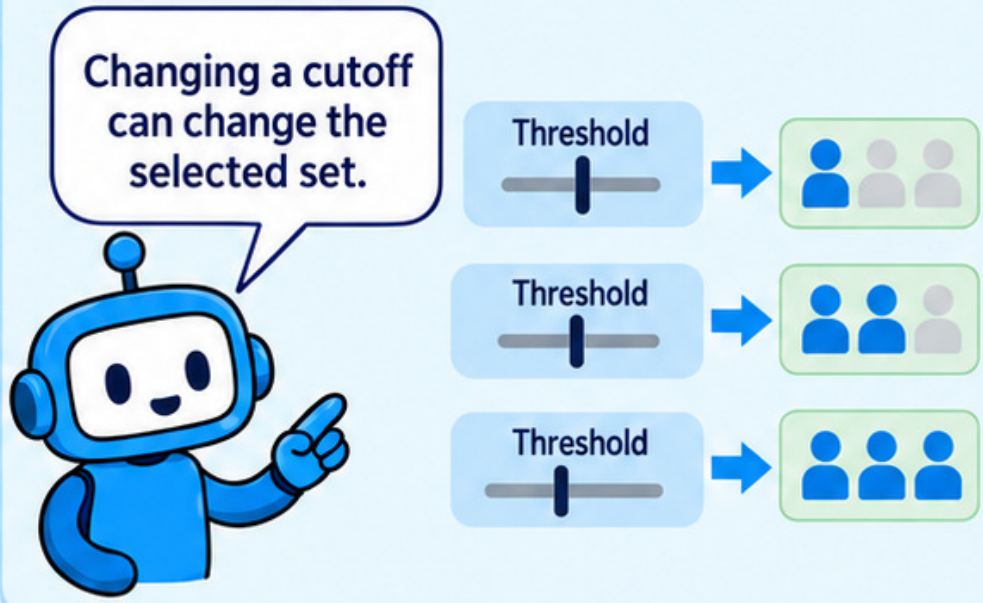




Check predicted probabilities against observed frequencies on suitable held-out data.



Changing the threshold changes selections; it does not repair a poorly calibrated probability model.



A SMALL WIN NEEDS MORE EVIDENCE

Repeat the comparison across relevant time periods and customer groups.

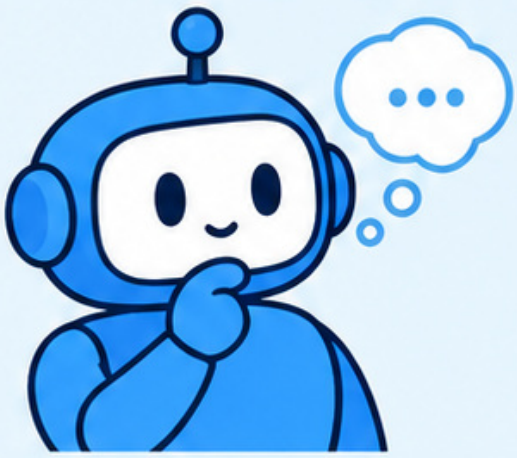
Time periods

Customer groups

Check mature-label coverage, review volume and missing-feature rates as well as precision and recall.

- Mature-label coverage
- Review volume
- Missing-feature rates
- Precision and recall

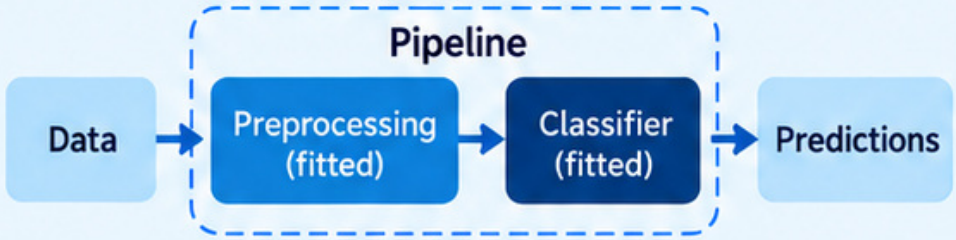
Eight examples cannot establish reliable performance for every customer. Recheck the chosen policy when the population or review capacity changes.



Time slice	Customer group	Cohort size	Label coverage	Queue volume	Precision	Recall

GIVE EACH TOOL A DEFINED JOB

scikit-learn Pipeline:
couple fitted preprocessing and classifier.

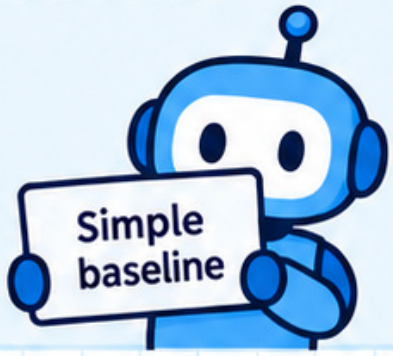


Metric functions:
count precision, recall and errors.

$$\text{Precision} = \frac{TP}{TP + FP}$$

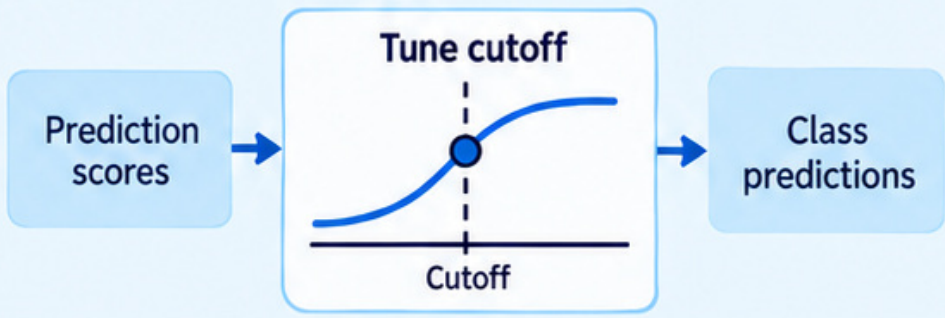
$$\text{Recall} = \frac{TP}{TP + FN}$$

DummyClassifier:
provide a simple classification reference.



- Most frequent
- Stratified
- Uniform

TunedThresholdClassifierCV:
tune a cutoff with a chosen scorer and split strategy.

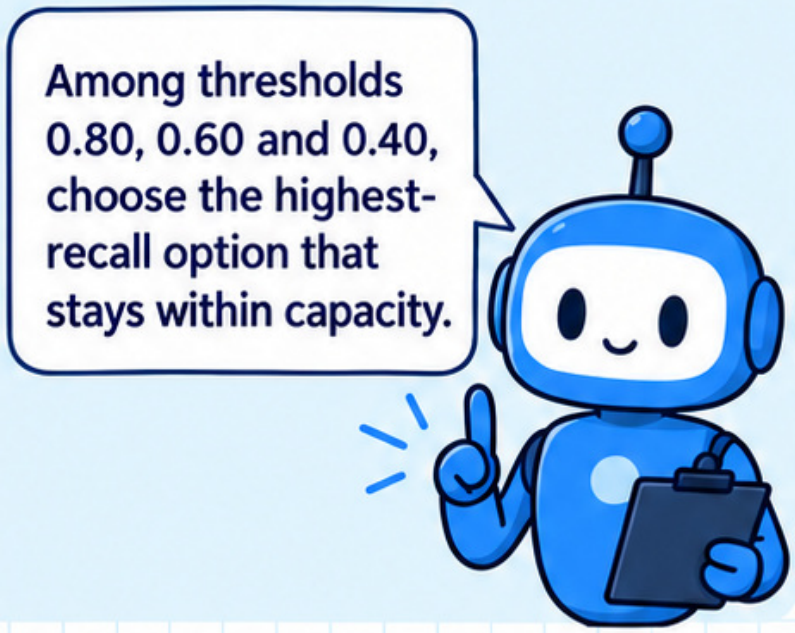


The application still defines capacity, tie rules and the final selection policy.

- Capacity
- Tie rules
- Final selection policy

YOUR TURN: CHOOSE UNDER A NEW BUDGET

Use the same toy threshold table.
Capacity is now five reviews.



List selected customers, true positives, false positives and false negatives. Calculate precision and recall.

Threshold:

Selected customers (IDs):

TP: FP: FN:

Precision: Recall:

Which dataset should guide this choice before the final test?

Dataset role:

RELEASE THE DECISION RULE WITH THE MODEL

Record the target definition, fitted pipeline, validation cohort, threshold or top-K rule, capacity policy, tie handling and measured results.

A model file alone does not describe what the application will do.

Artifacts

Model artifact

Fitted pipeline and model

Decision-policy artifact

Threshold or top-K rule, capacity policy, tie handling

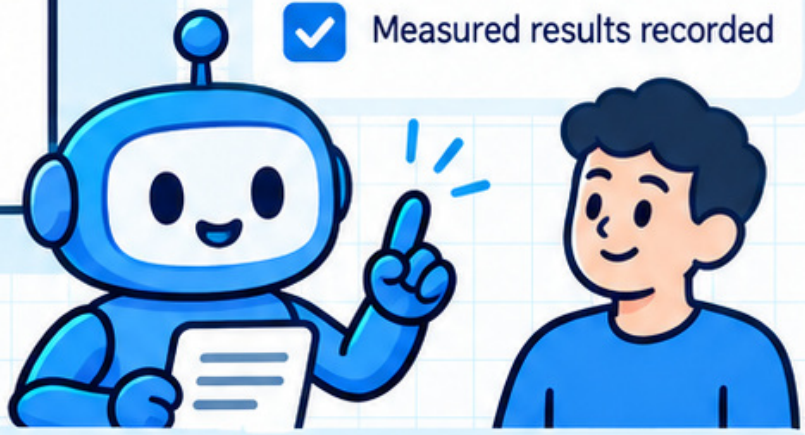
Versioned release manifest

Model Release v1

- Target definition
- Fitted pipeline
- Validation cohort
- Threshold or top-K rule
- Capacity policy
- Tie handling
- Measured results

Release checklist

- [x] Target definition recorded
- [x] Fitted pipeline recorded
- [x] Validation cohort recorded
- [x] Threshold or top-K rule recorded
- [x] Capacity policy recorded
- [x] Tie handling recorded
- [x] Measured results recorded



DAY 12

DAY 12: Track experiments and artifacts.
Save enough evidence to reconstruct the chosen system.