

MODEL SELECTION RATIONALE

Analyst: Tom Becker, AI Governance Lead (exemplar)
Organization: Corvana Bank
System: LimitWise
Evidence: Week 4 modeling notebook, sections 1 to 11, and the modeling team's results

Instructions

A model selection rationale is the document that records why one model was chosen over the others and at what point it will act. It is the artifact a regulator, an auditor, or your successor reads to understand the decision, so it is written from evidence, with the notebook section beside every number. It has five sections. The first is completed by the AI Governance Lead and inventories the evidence. You write sections 2 through 5: the champion and why, the operating point and the cost argument (the operating-point decision record), what the choice commits the program to at evaluation and in monitoring, and the open items and limitations the choice carries. Work top to bottom. Compare models on the cross-validation table, not on any single fold and not on the held-back set. Where the evidence does not settle a question, say so. Every number you write must be visible in the Week 4 notebook.

SECTION 1: THE EVIDENCE REVIEWED

Objective: Inventory the evidence the decision rests on and the conditions the Chief AI Officer's Week 3 memo put on the comparison, so that every reader starts from the same table.

Completed by the AI Governance Lead. Read it, but do not rewrite it.

What evidence did the modeling notebook produce, and in which section is each piece printed?

The Week 4 notebook runs on the working file the datasheet describes: 29,995 accounts, 27 model inputs, 22.1 percent default (section 1). It executes the split recorded in the remediation log, 23,996 training rows and 5,999 rows in the test set (section 2), fits every transformation on training rows only, 98 encoded columns (section 3), and trains three models on one fixed configuration each with balanced class weights (section 4). The evidence for the choice is the five-fold cross-validation table inside the training rows: mean, high fold, low fold, and spread for ROC-AUC, PR-AUC, and precision, recall, and F1 at the 0.50 reference threshold (section 5). Supporting it: ROC and precision-recall curves

from out-of-fold predictions (section 6), a threshold sweep from 0.20 to 0.80 (section 7), a capacity table at the top 5 to 25 percent of the ranked list (section 8), the test-set table read once (section 9), two notes for the file (section 10), and model_facts.json (section 11).

What did the Week 3 decision memo allow this comparison to be made on, and what did it rule out?

The Chief AI Officer's Week 3 memo opened the decision gate (her permission to compare models on this file) with two conditions.

First, models are compared on the cross-validation table, and the test set is read once.

Second, no probability threshold is set on this extract, because its 22.1 percent default rate is how the sample was drawn, not the rate on the bank's card book, which is near 3 percent. The use gate (i.e., her permission to contact a cardholder or touch a line on a score) stays closed until three things are true: the calibration check passes (predicted and observed default agree by score band on a holdout at the book rate, within 25 percent), the reason-coverage check passes (a plain-language reason exists for every value of every input, 100 percent), and the Week 2 watch list has run at the evaluation stage. The memo also carried forward the decision this document records: if the best-scoring model and the model whose reasons read like English are not the same model, this office brings her that choice, with the operating point and the cost of each error. The amount unit is still unknown, so no dollar figure taken from the amount fields appears in any cost statement (the no-dollar-figure rule).

Which success and failure criteria from the Use Case Assessment apply to this choice?

Two of the program's stated ambitions became the criteria this choice answers to.

Success: see trouble a statement cycle earlier than the behavioral scorecard does, which is what detection measures, and every adverse account action carries reasons the bank can put in a notice and defend to the Office of the Comptroller of the Currency, which is what explainability measures.

Failure: a repeat of the holiday cuts (i.e., line actions at scale on scores the bank was sure of and could not explain), and outreach that skews against a protected group, which the Week 1 memo's fourth condition measures monthly.

The Chief AI Officer's Week 1 decision set the posture: preventive first, meaning that in phase one the score leads only to outreach calls and hardship offers, a person reviews every contact, and no line is touched until validation and reason coverage are proven. A

champion that scores higher and cannot carry a reason fails the second success criterion before it reaches the first.

SECTION 2: THE CHAMPION AND WHY

Objective: Name the model that goes forward and defend the choice as a governance argument: performance read with its spread, weighed against interpretability, the organization's duty to explain, and the regulatory expectation in this domain.

Which model is the champion, and which fixed configuration?

The champion is the model that goes forward to the evaluation stage. It is logistic regression, in the configuration fixed before the first fit: regularization $C = 0.1$, scaled inputs, balanced class weights (section 4). Gradient boosting (15 leaves per tree, learning rate 0.05, 300 trees, section 4) has the higher mean ROC-AUC, 0.788 against 0.774 (section 5).

Logistic regression goes forward for two reasons. The gap between the two is no larger than one model's spread (i.e., the distance between its lowest and highest fold result), so the table does not show that gradient boosting is better. And the bank must be able to explain any adverse action, and a logistic regression score is a sum of fixed weights, which gives the same reason for the same facts.

Gradient boosting is the challenger: it is scored beside the champion in the Week 6 checks, which have not run yet, and it replaces the champion if those results say so. The decision tree is not retained. It trails both on every ranking metric by more than its spread (section 5).

What does the cross-validation table show, read with the spread? Where do the models differ by more than the spread, and where do they not?

A fold is one of the five pieces of the training set each model was scored on in turn. The spread is the distance from the lowest fold result to the highest. Mean ROC-AUC: 0.788, 0.774, and 0.741 for gradient boosting, logistic regression, and the tree, with spreads of 0.014, 0.018, and 0.010 (section 5). The gap between the first two, 0.014, equals gradient boosting's spread and is below logistic regression's, and the fold ranges, 0.778 to 0.792 and 0.763 to 0.780, overlap (section 5). On PR-AUC, 0.563 against 0.542 with spreads of 0.053 and 0.042, the ranges overlap (section 5). The table separates the tree from both, 0.033 behind logistic regression with ranges that do not touch. It does not separate the first two on any metric.

What does the choice gain or cost in interpretability and in the organization's ability to explain an outcome to the person affected? What does the domain's regulatory expectation require?

A logistic regression score is a sum: each input times a fixed weight. The reason for a score is the inputs that added most to it. So a sentence such as "the balance relative to the credit limit was higher than the portfolio's" is true of the account and of the model, and it stays true after a retrain as long as the weights stay close.

Gradient boosting builds its score from 300 trees. Two accounts with the same top input can be scored in opposite directions, and its reasons have to be estimated afterward by an attribution method (i.e., a separate calculation that assigns credit for the score to each input).

Regulation B, 12 CFR 1002.9, requires an adverse-action notice to state the specific principal reasons, and a line decrease is an adverse action. The reason-coverage check (i.e., does a plain-language reason exist for every value of every input) has not run yet for either model. The cost of choosing logistic regression: 0.014 of mean ROC-AUC (section 5), and at the top 10 percent of the ranked list, 27 fewer defaulters caught out of 2,400 flagged accounts, 1,666 against 1,693 (section 8).

What evidence would change this choice?

Three results would change it. None is in yet.

First, the reason-stability check runs (i.e., the modeling team retrains the model on two random halves of the training set and compares the top reason for every flagged account). If logistic regression's top reasons change for more than 10 percent of flagged accounts and gradient boosting's do not, the main argument for logistic regression is gone.

Second, the calibration check runs and the probabilities are corrected to the book rate. If gradient boosting still catches more defaulters than logistic regression in the top 10 percent after that correction, and the difference is larger than the fold-to-fold spread, the gap is real and not noise. On the extract, the difference is 27 defaulters out of 2,400 flagged accounts (section 8).

Third, the program stays in phase one (i.e., outreach calls and hardship offers only, no line changes), so no adverse-action notice is needed, and gradient boosting passes the reason-coverage check. Then the duty to explain no longer favors logistic regression. A better number in the test-set table would not change the choice. That table is read once and is not used to choose.

SECTION 3: THE OPERATING POINT AND THE COST ARGUMENT

Objective: Record the operating-point decision: where the model acts, what the notebook shows at that point, what each kind of error costs this organization, and what has to be true before the point is used. This section is the operating-point decision record.

Where is the operating point: a share of the ranked list or a probability threshold? What does the notebook show at that point (flagged, caught, missed, false alarms, precision, recall), and in which section?

The operating point is the place in the ranked list where the model acts. It is a share of the list, not a probability, and provisional until recalibration to the book rate: the top 10 percent, the group the Week 2 memo defined as the one LimitWise would act on. Section 8 shows the champion at that share flagging 2,400 of the 23,996 training rows, catching 1,666 defaulters, missing 3,642 (misses are defaulters not flagged), with 734 false alarms (i.e., paying accounts flagged), precision 0.694 (i.e., flagged accounts that defaulted), recall 0.314 (i.e., defaulters caught), implied threshold 0.821. The challenger's row: 1,693 caught, 3,615 missed, 707 false alarms (section 8). Every count is an extract count. On a book near 3 percent default, the counts change and precision falls. The share does not.

What does each kind of error cost this organization, in its own terms, and which one hurts more?

A missed defaulter (i.e., a defaulter the model did not flag) charges off without the call that might have placed the account on a hardship plan, and the Chief Credit Officer's office counts it. A false alarm (i.e., a paying account the model flagged) costs different things in different phases. In phase one (outreach calls and hardship offers only) it costs an outreach call to a customer who was going to pay, and a reviewer's time.

In any later phase, when a score can lead to a line decrease, it is a line decrease on a paying customer. That is an adverse action under Regulation B, and in this portfolio one point of false alarm rate is 46,000 people (Week 2 memo). So in phase one the miss hurts more, and the operating point leans toward catching more defaulters, as far as outreach capacity allows. In any phase that touches a line, the false alarm hurts more, and the point moves the other way. No dollar figure is attached to either error, because the amount unit is unknown and the no-dollar-figure rule applies.

What has to be true before this operating point is used on a real person? Which use-gate conditions from the Week 3 memo does that include?

The use gate must open. The Chief AI Officer's Week 3 memo sets its three conditions:

- the calibration check,
- the reason-coverage check, and
- the Week 2 watch list run at the evaluation stage.

These conditions are not repeated here. None has run yet.

This rationale adds three conditions of its own.

- The share is confirmed against stated outreach capacity, which the program file does not yet hold.
- A line action needs a second decision, independent validation, and a notice whose reasons have passed the reason-coverage check and the reason-stability check.
- No account in a declared hardship is actioned on the score.

None of these conditions is met today.

Who owns the operating point, and when and how is it revisited?

The AI Governance Office holds the operating-point record, and the Chief AI Officer approves any change. The modeling team supplies the capacity table at every retrain. In Week 6, when the calibration check reports: if calibration holds, the share is re-read on the book-rate holdout and confirmed or moved, with the row quoted. At each monthly cycle: outreach pauses and the point is reviewed if the flagged share drifts from the set share, if the group false-alarm check (the false alarm rate by group) finds a group of 1,000 or more more than 20 percent above the portfolio false alarm rate, or if any score band fails the 25 percent calibration tripwire. Before any phase that touches a line: the share is set again from scratch. Customer Advocacy, the unit named in the organization briefing, reviews contact rates by group before they are filed.

SECTION 4: WHAT THE CHOICE COMMITS THE PROGRAM TO

Objective: Carry the choice forward: the champion is what the evaluation stage audits and the operating point is what monitoring watches, so name the tests and tripwires now.

Which watch-list tests from the Chief AI Officer's memos must the champion pass at the evaluation stage, and what are their tripwires?

The checks are the six watch-list items in the Chief AI Officer's Week 2 memo, each with its tripwire, restated as controls in her Week 3 memo with the STATUS-versus-PAID check added. This rationale adopts all seven without change and does not repeat them. None has

run yet. One of them now carries the weight of this choice: the reason-stability check (Week 2 memo, item six). Logistic regression was chosen for fixed, stable reasons. If the champion fails that check, the argument in section 2 is gone and the challenger is read again. The calibration check is explained at the evaluation stage, not here.

What will be monitored in operation, against what tripwires, and who owns each?

Monitoring follows the Week 1 memo's fourth condition (error and contact rates by group, monthly) and the Week 2 watch list, which names an owner for each item. This rationale adds one item. Because the operating point is a share of the ranked list, the flagged share is monitored each cycle against the set share and against outreach capacity, which the program file does not yet state. The program owns it. If the share drifts, the point is reviewed. The reason-stability check is re-run at every retrain rather than monthly, because retraining is what moves the reasons.

What does the held-back table (read once) show, and why is it not the basis for the choice?

The test set is the 5,999 rows split off in Week 3 (section 2). Section 9 scores it once, at the 0.50 reference threshold. Logistic regression: ROC-AUC 0.772, PR-AUC 0.530, precision 0.481, recall 0.570, F1 0.522. Decision tree: 0.736, 0.466, 0.532, 0.480, 0.505. Gradient boosting: 0.782, 0.556, 0.468, 0.621, 0.534 (section 9). Every ROC-AUC sits inside its model's fold range from section 5, in the same order. The table says the cross-validation estimate was honest. It is not the basis for the choice, for two reasons. It is one number per model with no spread, so it cannot say whether a difference is real. And a set used while choosing is no longer a test set, so Week 6 would have nothing left to confirm the champion on.

SECTION 5: OPEN ITEMS AND LIMITATIONS

Objective: Record what the choice inherits from the datasheet and what it adds, so the next reader knows what this model is not allowed to be used for and what is still unknown.

Which open items from the datasheet constrain this model, and how?

The open items are the nine in section 5 of the datasheet. The Chief AI Officer's Week 3 memo closed three and left six open with a control each. They are not repeated here. Three constrain the champion directly.

1. The sampling method: no probability this model prints is a book probability, so the operating point is a share and no threshold is set.
2. Status codes 0 and -2: the six strongest inputs carry a level that reads "no delinquency recorded, code undefined," so no reason built on it reaches a cardholder until the code is defined.
3. The amount unit: no dollar figure in any cost statement or notice.

The other three (the month rule, the supplementary card, the label) constrain the evaluation and the loss forecast, as the Week 3 memo states.

What limitation does this choice add that the datasheet did not carry?

Two. First, the champion was chosen for a property it has not yet demonstrated. The six balance fields it weighs correlate at 0.80 to 0.95 (Week 2 memo), so weight can move between near-identical fields at each retrain. The reason-stability check has not run yet, and this rationale does not assume the champion passes it. Second, the choice gives up detection: 0.014 of mean ROC-AUC (section 5) and 27 accounts at the top 10 percent of the extract (section 8). On the book that gap is unmeasured. A smaller matter: the pipeline fills empty PAY_RATIO_RECENT values with the training-set median (section 3). That fill is a fitted step production has to reproduce, or the model serves on a field it never saw empty.

What minority position was argued, and why was it not adopted (or what would make it the right call)?

The Chief AI Officer's Week 3 memo names it: shadow mode first (score the live book with no action attached, calibrate on it, then choose). It is the better experiment, because the gap between the models is measured on an enriched extract and the book is where the choice is judged. It is not adopted, for two reasons. The Week 6 checks are written against a named champion, so the program needs one now. A shadow run scores real cardholders while the use gate is closed, which needs its own ruling. What stays: gradient boosting runs beside the champion through the calibration, merged-codes, and reason-stability checks. If the calibration check fails for both, shadow mode becomes the plan.