MACHINE LEARNING · CREDIT RISK & RESPONSIBLE AI

Credit Default Risk Model: *the threshold is the policy*.

An end-to-end supervised-ML credit model that separates ranking default risk from the lending decision, then makes the cost of the cutoff visible, alongside the checks that say whether to trust the model: held-out performance, calibration, feature importance, and a fairness audit.

Author S. Ize-Iyamu **Audience** Credit risk & ML teams **Length** 2 pages **Status** Technical brief

Stack scikit-learn · fairlearn · Streamlit

The Problem

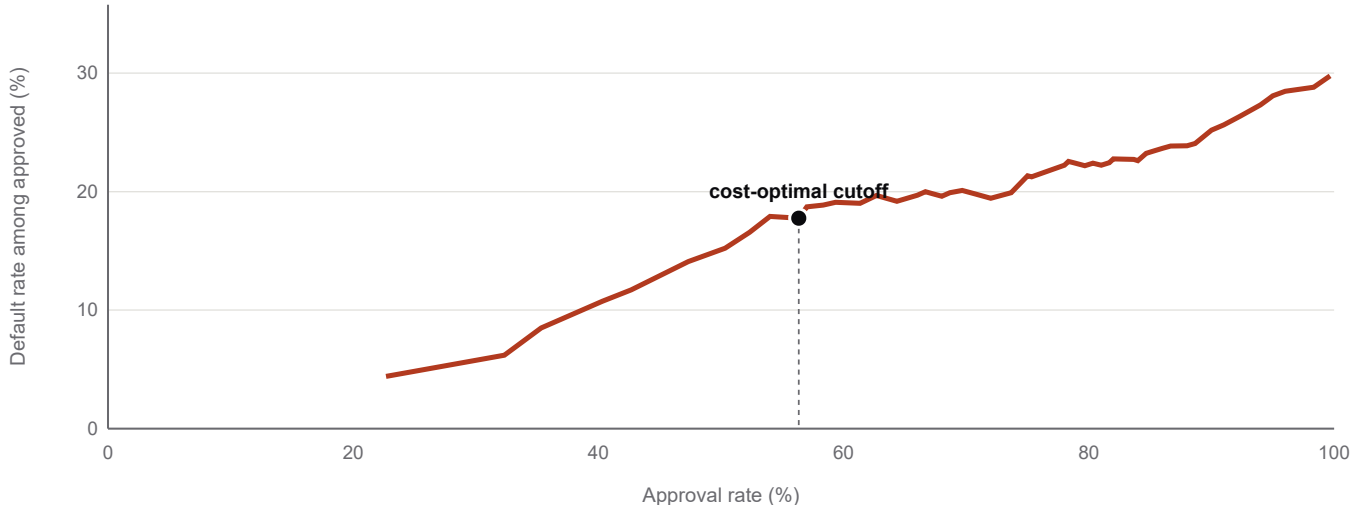
Approving an applicant is a policy decision, and the model only supplies the input to it. It ranks default risk as a probability. Turning that probability into approve or decline takes a threshold, and the threshold is a business call about the cost of a default against the value of an approval, which the model can't make on its own. The Statlog German Credit data ships with an asymmetric cost matrix, where scoring a bad applicant as good costs five times the reverse, so the cutoff is what's really being set. This model surfaces that trade-off and the checks behind it.

BOOSTING AUC	BASELINE AUC	TEST SET	COST MATRIX	AUDIT
0.76	0.80	300 held out	5 : 1	fairlearn

The Threshold Trade-Off

The model outputs a default probability, and a separate, user-set threshold turns it into approve or decline. A tight cutoff approves fewer people and keeps the book clean; loosening it brings in more applicants and a higher default rate among them. The chart traces that operating curve on the held-out test set. The 5:1 cost matrix implies a cost-optimal cutoff near a one-in-six default probability, about 56% approved here, the marked operating point used for the fairness audit.

FIGURE 1 · THE OPERATING CURVE



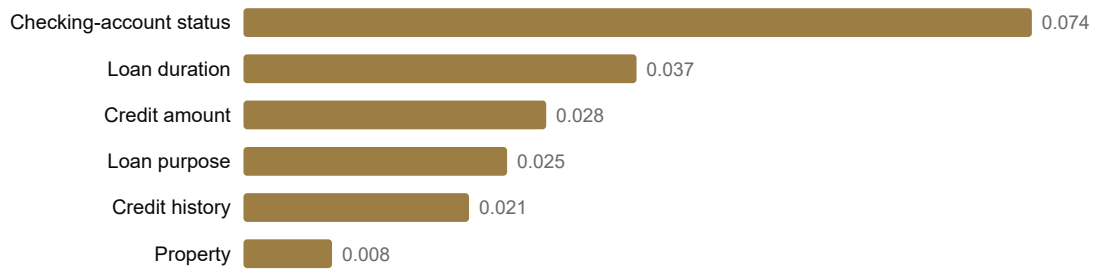
As the policy widens, the default rate inside the approved book climbs from near zero toward the portfolio base rate. The marked point is the cost-optimal cutoff set by the 5:1 cost matrix; a lender picks where on this curve to sit.

> THE POINT
>
> What makes a credit model interesting is the threshold you set and what that costs, more than the AUC. Conflating the model with the decision is the most common mistake in applied credit modeling, so this build keeps them separate by design.

Is the Added Complexity Worth It?

Gradient boosting is reported against a logistic-regression baseline, the honesty check. On structured credit data of this size the baseline is competitive and here slightly ahead (**AUC 0.80 against 0.76**), a reminder to check that the added complexity is worth keeping. Preprocessing is fit only on the training fold, so the scaler and encoder never touch the test set, the choice that decides whether a reported AUC is real.

FIGURE 2 · PERMUTATION IMPORTANCE (DROP IN AUC WHEN SHUFFLED)

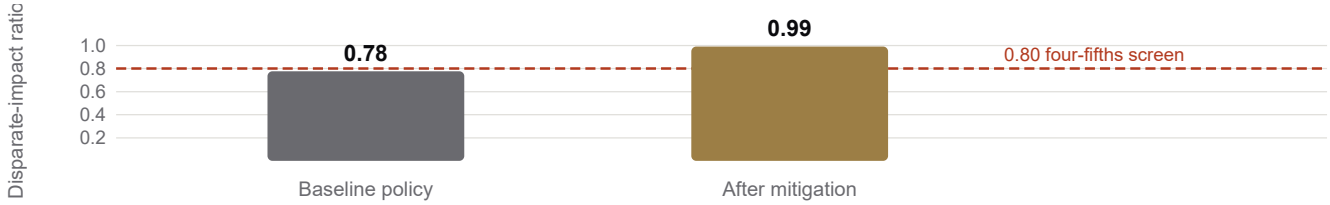


Checking-account status carries the most signal, which is what the credit literature expects.

Responsible-AI Audit

The data carries attributes correlated with protected characteristics, so a fairness pass belongs in the build. The audit reframes to the favorable outcome a person experiences, **approval**, and measures parity by age group (under 25 against 25 and older). At the cost-optimal cutoff the model just misses the four-fifths screen (ratio 0.78): younger applicants are approved at a lower rate. Group-aware thresholds calibrated to equalize approval rates lift the ratio to parity (0.99), holding overall approval and book quality flat by reallocating approvals across groups rather than approving more people.

FIGURE 3 · DISPARATE-IMPACT RATIO, BEFORE AND AFTER MITIGATION



A ratio below 0.80 (US EEOC four-fifths rule) flags disparate impact; the baseline falls just under it, and equalizing approval rates by group lifts it to parity.

METRIC	BASELINE POLICY	AFTER MITIGATION
Disparate-impact ratio (approval)	0.78	0.99
Equalized-odds gap	0.10	0.07
Approval rate	56%	56%
Default rate among approved	18%	18%

The book-quality cost is near zero, but there's still a catch: the mitigation uses a protected attribute directly in the decision, which can be unlawful in real lending. Treat it as analysis of the trade-off rather than a deployment recipe.

Limitations & Sources

Honest scope. The German Credit set is 1,000 rows, so estimates are noisy, with no out-of-time validation or loss-given-default economics; a deployed scorecard would need these plus legal review of any group-aware step. **Method:** a scikit-learn pipeline, `HistGradientBoostingClassifier` against a `LogisticRegression` baseline, stratified hold-out, permutation importance, and a fairlearn audit; the cutoff is the cost-optimal point from the 5:1 cost matrix and the mitigation a group-aware threshold calibration. Data: Statlog German Credit (Hofmann, 1994) via OpenML `credit-g`; four-fifths rule from US EEOC (1978). Live app at credit-default-risk-model.streamlit.app.