Cost-Sensitive Credit Card Approval Policy

Deliverable: an operational decision policy with explicit risk trade-offs

Marco Frova

Bending Spoons

February 6, 2026

Executive summary

- **Stakeholder question:** “Which applicants should we approve to *maximize expected value* given asymmetric costs of false approvals and false rejections?”
- **What I did:** built an interpretable *probabilistic* credit risk model, calibrated predicted default probabilities, and optimized approval decisions using a cost-sensitive threshold with a manual review band.
- **Key finding:** models with similar classification performance (e.g. AUC) led to *very different expected costs* once probability calibration and decision thresholds were considered.
- **Why it matters:** using accuracy-driven models or a naive 0.5 cutoff produces **overconfident approvals** and systematically underestimates financial risk.
- **Recommendation:** deploy a calibrated model, choose thresholds by **expected cost minimization**, and route borderline cases to manual review; monitor calibration and cost over time.

Problem framing

Context. A credit card issuer receives a stream of applications. For each applicant we observe a set of features (income, credit history, utilization proxies, etc.) and we want to decide whether to approve or decline.

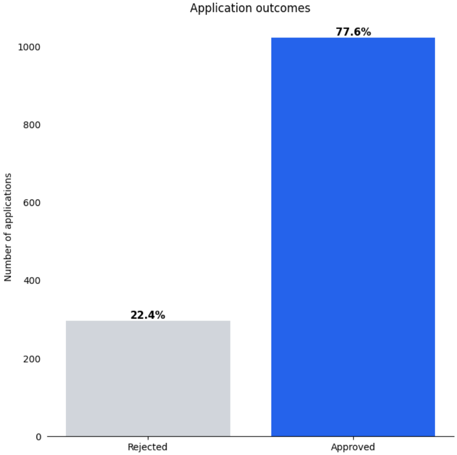
Decision need. The stakeholder needs an approval policy that:

- produces a **risk score as a calibrated probability** of approval;
- translates risk into **actionable rules** (approve / decline / manual review);
- explicitly accounts for **asymmetric error costs** (false approvals are typically more expensive than false rejections);
- respects **operational constraints** (limited manual review capacity).

What success looks like. Minimum **expected decision cost** under realistic assumptions, with **transparent** thresholds and **well-calibrated** probabilities.

Data

- **Dataset:** 1,319 credit card applications (features available at application time).
- **Target:** `card` = approved vs rejected (*proxy* for historical approval, not default/profit).
- **Features:** credit history (`reports`, `active`), capacity (`income`, `owner`), stability (`age`, `months`, `selfemp`), exposure (`majorcards`).
- **Prep:** stratified 80/20 train–test; standardize continuous features on train stats.
- **Base rate:** approvals ≈77.6% ⇒ cost-sensitive thresholds + manual review band.



Approach

Core idea: turn applicant features into a **calibrated probability** of approval, then convert probabilities into an **operational policy** under asymmetric costs.

How we model risk.

- **Logistic regression** for robustness and interpretability (coefficients / odds ratios).
- Trained via **maximum likelihood** (IRLS); continuous features **standardized** on train statistics.

How we evaluate predictions.

- 80/20 stratified train–test split; no leakage in preprocessing.
- Metrics for ranking (ROC AUC, PR AUC) and probability accuracy (Brier score, calibration curve).

From model to decisions.

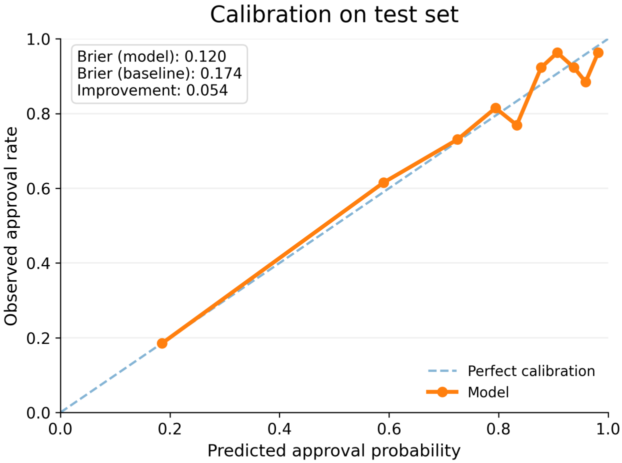
- Choose approval threshold(s) by **minimizing expected cost** with $\text{FP} \gg \text{FN}$.
- Add a **manual review band** for borderline probabilities (capacity constraint).

Performance (test set)

- **Setup:** held-out **20%** test set; base approval rate ≈ 77% ⇒ accuracy alone is misleading.
- **Probability accuracy:** Brier = **0.1200** vs baseline (predict train prevalence) **0.1735** ⇒ improvement **0.0536**.
- **Calibration:** predicted probabilities align well with observed rates; good enough for thresholding + review band.
- **Ranking:** ROC AUC = **0.8104**; PR AUC = **0.9146**.

Probability accuracy & calibration

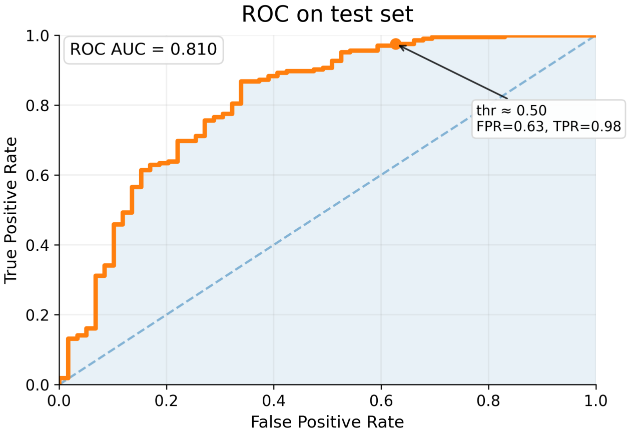
- **Brier:** 0.1200 (baseline 0.1735) ⇒ probabilities are informative.
- **Interpretation:** observed approval rates are close to predictions; small deviation only at the extreme upper tail.



Ranking quality & robustness

- **Ranking (test):** ROC AUC = **0.8104**; PR AUC = **0.9146**.
- **Stability (20 splits):** ROC AUC mean **0.8316** (sd 0.0349); Brier mean **0.1148** (sd 0.0098).
- **Default threshold issue:** at $t = 0.50$ recall **0.971** but specificity **0.373** ⇒ many false approvals.

Metric	**Mean**	**P10**	**P90**
ROC AUC	0.832	0.796	0.876
Brier	0.115	0.104	0.128



Drivers & interpretability

- **Interpretable model:** standardized features ⇒ each coefficient is a **1 SD effect** on log-odds.
- **Takeaway:** approvals are primarily driven by **credit history** (`reports`) and **credit activity** (`active`).
- **Policy intuition:** even small increases in `reports` require strong compensating signals (e.g., `active`, `income`) to reach high p(approve).

Driver	**OR**	**Dir**	**Meaning**
`reports`	0.09	↓	Dominant negative signal: derogatory history sharply reduces approval odds
`active`	2.20	↑	Strong positive signal: active credit profile boosts approval likelihood
`income`	1.50	↑	Higher income increases perceived repayment capacity
`owner`	1.35	↑	Home ownership acts as a stability proxy
`dependents`	0.67	↓	More dependents increase financial burden and reduce approval likelihood

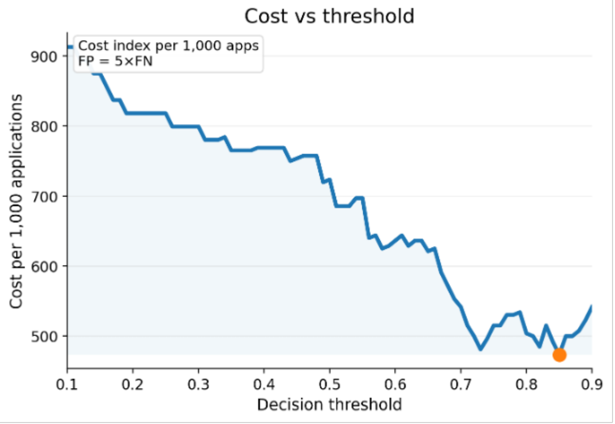
Robustness & policy implications

- **Statistical robustness:** Wald CIs from IRLS show main drivers (`reports`, `active`, `income`) are stable (effects do not cross 0).
- **Estimation robustness:** Bayesian logistic regression with weakly-informative Gaussian prior yields similar posterior means and credible intervals $\Rightarrow$ conclusions are not framework-sensitive.
- **Implication for policy:** because `reports` is dominant, even modest negative history requires strong compensating signals to reach high $p(\text{approve})$.
- **Operational takeaway:** under FP$\gg$FN, a conservative threshold is justified; ambiguous mid-scores should go to **stricter thresholds or manual review band**.

Policy trade-off & threshold selection

- **Goal:** turn scores into an operational approve/decline policy.
- **Cost framing:** Cost $= C_{FP} \cdot FP + C_{FN} \cdot FN$.
- **Reference scenario:** $C_{FP} = 5 \times C_{FN}$ (cost index per 1,000 apps).
- **Result:** cost is minimized at a **high threshold** ($t^* = 0.85$), prioritizing fewer false approvals.

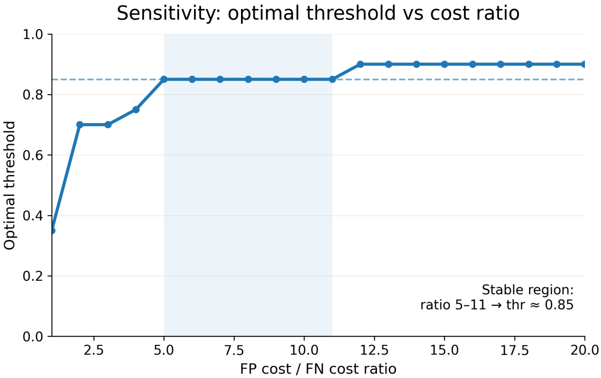
Metric at $t^* = 0.85$	**Value**
Approval rate	50.8%
Precision	93.3%
Recall	61.0%
Specificity	84.8%



Sensitivity analysis: threshold vs FP/FN cost ratio

- **Why:** FP/FN costs are uncertain and can vary across time/orgs.
- **Method:** recompute the **cost-minimizing threshold** as FP/FN varies (1 to 20).
- **Key result:** **plateau** for FP/FN ≈ 5–11 ⇒ recommended threshold is **robust**: $t^* \approx 0.85$.

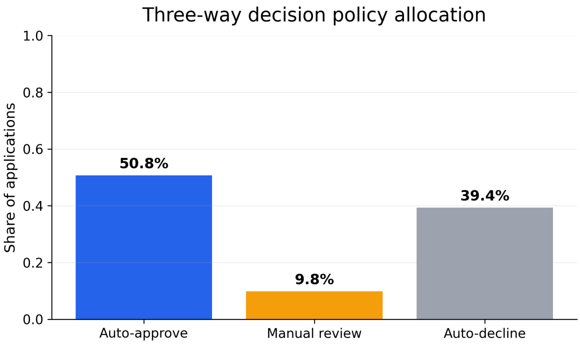
FP/FN	t^*	Approval rate	Cost / 1,000
2:1	0.70	78.0%	268.9
5:1	0.85	50.8%	473.5
10:1	0.85	50.8%	643.9



Operationalization: three-way decision policy

- **Why:** borderline cases need human review (capacity constraint).
- **Bands (from calibrated p(approve)):**
  - Auto-approve: $p \geq t_{\text{high}} = 0.85$
  - Manual review: next $\approx 10\%$ below t_{high}
  - Auto-decline: remaining low-probability cases
- **Allocation (test):** 50.8% approve, 9.8% review, 39.4% decline.

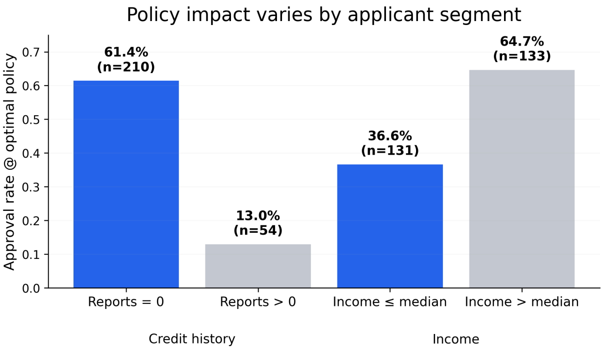
Manual handling	**FP**	**FN**
Conservative (treat as decline)	9	80
Aggressive (treat as approve)	15	60



Segment insights (policy behavior at $t^* = 0.85$)

- **Credit history is a gate:** any derogatory `reports` shifts policy to a highly conservative regime.
- **Approval volume differs sharply:** `reports=0` → **61.4%** approved (n=210) vs `reports>0` → **13.0%** (n=54).
- **Income acts as a positive tilt: 36.6%** approved (income $\leq$ median, n=131) vs **64.7%** (income $>$ median, n=133).

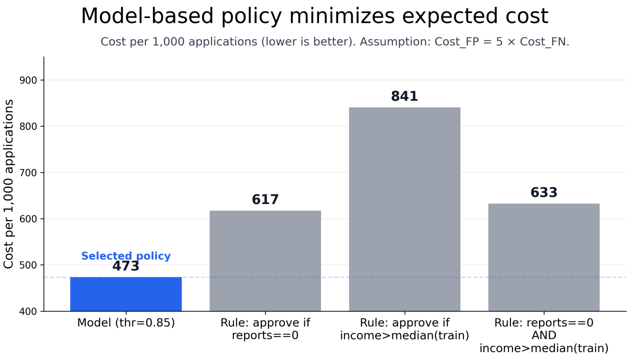
Implication: keep the global threshold for simplicity, but consider **segment-specific manual review rules** for `reports>0` applicants.



Rules-based benchmark (test set, FP = 5×FN)

- **Goal:** verify value beyond simple heuristics.
- **Result:** model-based policy is **lowest cost**: **473** per 1,000 vs rules **617–841**.
- **Why rules lose:** high-volume rules ⇒ many **false approvals** (expensive); conservative rules ⇒ many **false rejections**.

Gain vs best rule: $617 - 473 =$ **144** cost units / 1,000 apps (≈ **23%** reduction).


Model-based policy minimizes expected cost

Cost per 1,000 applications (lower is better). Assumption: Cost_FP = 5 × Cost_FN.

Limitations

- **Proxy target:** `card` is historical approval, not default / profit ⇒ optimizes approval propensity, not true value.
- **Costs are scenario-based:** FP/FN costs are an *index* (not $) ⇒ need calibration from losses, recovery, CLV.
- **Validation is not time-aware:** random splits miss drift / macro shifts ⇒ require time-based CV + monitoring.
- **Fairness & regulation:** no explicit constraints; segment outcomes differ ⇒ needs compliance checks + constraints if required.
- **Manual review modeled by bounds:** no reviewer outcomes ⇒ planning tool, not measured operational performance.

Next steps & takeaway

- **Move to business outcomes:** model default / loss / profit; optimize expected value.
- **Monetize the cost function:** estimate FP/FN in $ using historical loss + revenue.
- **Deploy safely:** time-based validation, drift + calibration monitoring, periodic recalibration.
- **Operational refinement:** learn from manual-review outcomes; tune review band + routing rules.
- **Governance:** fairness monitoring and (if needed) segment-aware constraints/policies.

Final takeaway: value is not accuracy alone — it's **calibrated probabilities + cost-aware policy** that is **interpretable, robust, and operationalizable**.

Appendix: Frequentist robustness check (Wald 95% CIs)

Goal: assess statistical stability of model coefficients under MLE (IRLS / Fisher scoring).

Var	$\hat{\beta}$	SE	95% CI$_\beta$	OR	95% CI$_{\text{OR}}$
reports	-2.432	0.223	[-2.870, -1.995]	0.088	[0.057, 0.136]
age	-0.105	0.112	[-0.324, 0.114]	0.900	[0.723, 1.121]
income	0.395	0.123	[0.154, 0.635]	1.484	[1.166, 1.887]
owner	0.306	0.114	[0.083, 0.529]	1.358	[1.087, 1.698]
selfemp	-0.141	0.084	[-0.306, 0.025]	0.869	[0.736, 1.025]
dependents	-0.397	0.095	[-0.584, -0.210]	0.672	[0.557, 0.810]
months	0.010	0.103	[-0.191, 0.212]	1.010	[0.826, 1.236]
majorcards	0.256	0.081	[0.097, 0.415]	1.292	[1.101, 1.515]
active	0.797	0.133	[0.536, 1.057]	2.218	[1.708, 2.879]

- **Key drivers** (reports, active, income) show tight intervals and stable signs.
- Secondary variables have weaker or non-significant effects, consistent with ranking results.

Appendix: Bayesian robustness check (RW Metropolis–Hastings)

Goal: verify that key drivers are stable under a Bayesian formulation (weakly-informative Gaussian prior).

Var	**MLE 95% CI**	**Bayes 95% CrI**	$\mathrm{OR}_{\mathrm{MLE}}$	$\mathrm{OR}_{\mathrm{Bayes}}$
reports	[-2.870, -1.995]	[-2.934, -2.023]	0.088	0.084
income	[0.154, 0.635]	[0.172, 0.648]	1.484	1.498
active	[0.536, 1.057]	[0.543, 1.064]	2.218	2.228

- **Conclusion:** posterior intervals closely match MLE confidence intervals $\Rightarrow$ same signs and magnitudes for the main drivers.