# SMB Credit: PD, LGD & Expected-Loss Pricing — Results & Recommendations

github.com/guptagaurav-info/smb-credit-pd-lgd · 772,591 real SBA 7(a)/504 loans (US SBA FOIA) · train FY1990–2003 · validate FY2004–05 · **test FY2006–07 (crisis cohort, default rate 38.8%)**

**0.664**
Champion GBM AUC (out of time) vs 0.616 scorecard

**0.964 → 0.664**
AUC with leaky Term field vs honest model

**31% vs 45%**
Bad rate of swap-ins vs swap-outs at 42.5% approval

**+$21M**
Illustrative lifetime profit vs scorecard at optimal cutoffs (+3.5%)

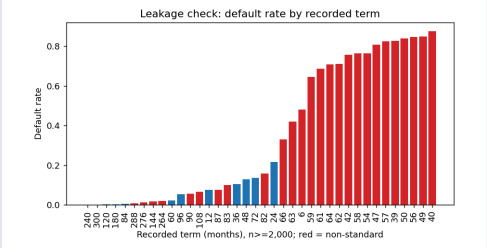
## FINAL OUTPUTS

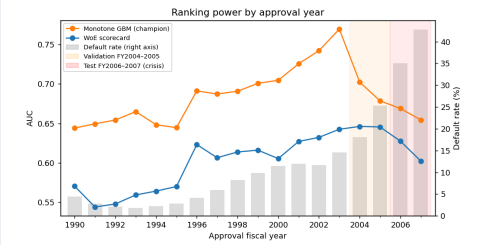
| Model (test FY2006–07) | AUC | KS | Brier |
|---|---|---|---|
| WoE logistic scorecard | 0.616 | 0.19 | 0.278 |
| **Monotone LightGBM (champion)** | **0.664** | **0.24** | **0.237** |

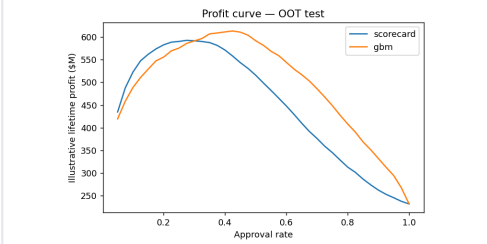
| Calibration (test) | Mean PD | ECE |
|---|---|---|
| Isotonic on FY2004–05 | 26.1% | 0.128 |
| Stress overlay (1.5× validation rate) | 32.8% | 0.077 |

| PD decile | Observed default | Indicated rate |
|---|---|---|
| 1 (lowest risk) | 9.1% | 8.8% |
| 3 | 29.7% | 10.3% |
| 5 | 39.6% | 10.9% |
| 8 | 52.4% | 12.2% |
| 10 (highest risk) | 54.5% | 13.6% |

Indicated rate = cost of funds 4.5% + opex 1.0% + capital 1.2% + margin 2.0% + annual expected loss. Illustrative economics.




## DATA & METHODOLOGY

### SOURCES

- US SBA 7(a) & 504 loan file (FOIA, public domain): 899,164 loans; mirror huggingface.co/datasets/MaddRaf/SBAnational; sha256-verified.
- Labels: charged off (CHGOFF) vs paid in full; charged-off principal for loss severity.
- Scope: FY1990–2007 resolved loans (772,591). Pre-1990 is survivor-biased; post-2007 immature.

### METHOD

- Raw lands as text; typed and cleaned in DuckDB SQL with a data-quality count per rule (2-digit-year fix, coded flags).
- Point-in-time network features: lender / industry / state charge-off rates using only outcomes known before each loan's fiscal year, shrunk toward and divided by the network rate.
- Out-of-time split by approval year; leakage test on suspicious fields.
- Models: WoE logistic scorecard vs LightGBM with monotone constraints, early stopping and isotonic calibration on validation; LGD champion chosen on validation with a 2% materiality bar.

### KEY CALCULATIONS

- AUC / KS: rank-ordering; Brier and ECE: calibration error by decile.
- Stress overlay: shift log-odds so mean PD = 1.5 × validation default rate (ranking unchanged).
- LGD lender = charged-off principal ÷ disbursed × (1 − SBA guarantee %); EL = PD × LGD × EAD.
- Rate = CoF 4.5% + opex 1% + capital 1.2% + margin 2% + EL ÷ 3-yr average life.
- Profit = good loans × 3.2% spread × 3 yrs − lender loss on defaults; cutoff = profit maximum; swap-set at that approval rate. PSI on 10 score bins.

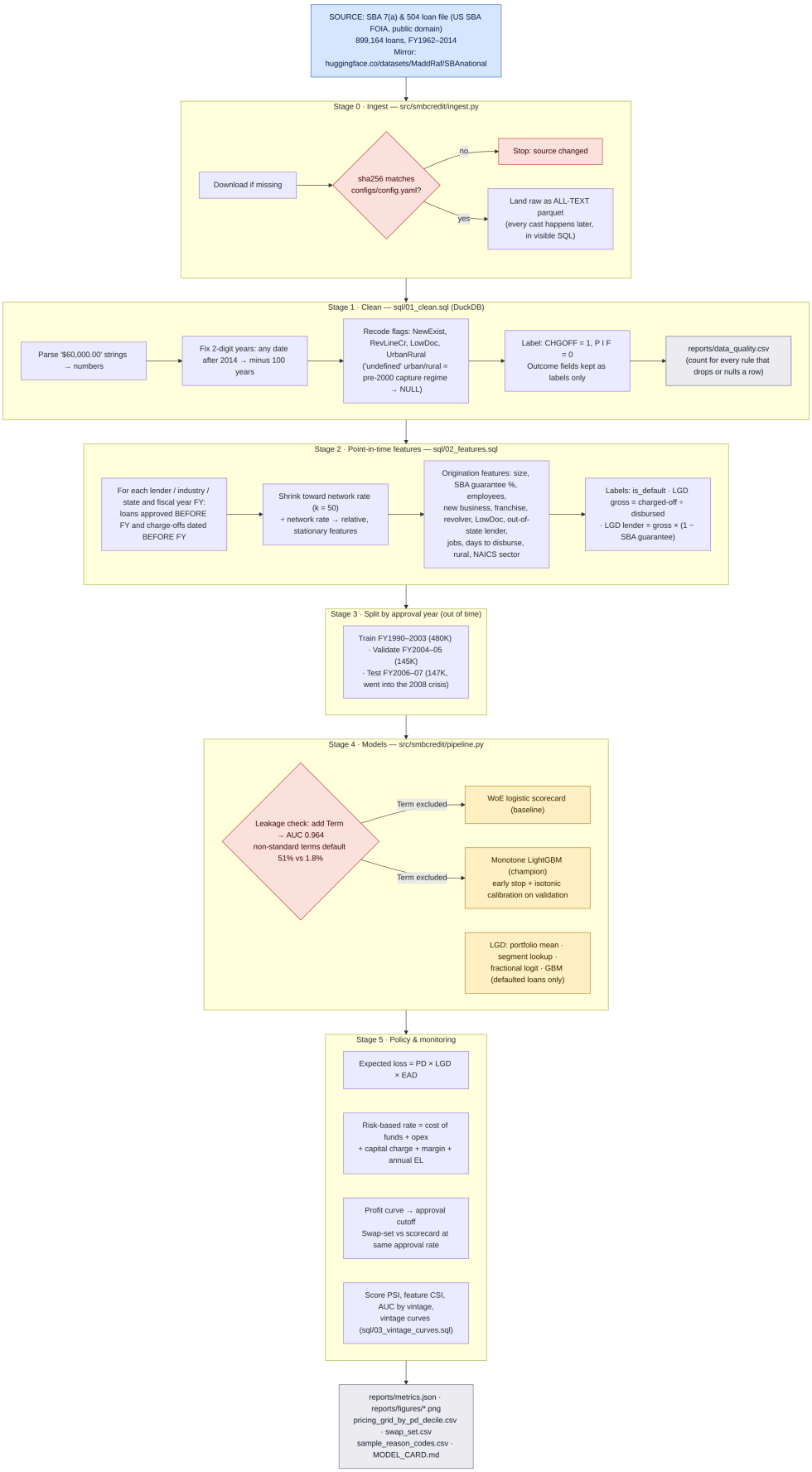
## RECOMMENDATIONS

1. **Adopt the monotone GBM as champion; keep the WoE scorecard as challenger.** It ranks better out of time and in every vintage, and monotone constraints plus TreeSHAP reason codes keep it explainable for adverse action.
2. **Ban the Term field and any post-origination fields by data rule.** Term carried post-default information (non-standard terms 51% bad vs 1.8%); add a leakage test to model validation.
3. **Separate ranking from level.** Use the stress overlay for loss forecasting and pricing in a downturn; recalibrate the intercept when the portfolio default rate moves, without retraining.
4. **Use portfolio-mean LGD × (1 − SBA guarantee) for now.** No LGD model beat the mean by the 2% bar; collect collateral, lien position and cash flow at default before building one.
5. **Price from the PD-decile grid and set the cutoff on profit, not AUC.** The swap-set shows where the gain comes from: same approval rate, materially worse loans declined.
6. **Do not deploy before retraining on recent vintages.** Score PSI test vs train is 1.64 (program mix shifted: guarantee % fell from ~85% to ~59%), which would trigger a recalibration review.
7. **Next step: add bank-transaction cash-flow features.** Origination fields cap AUC near 0.66; the companion cash-flow project reaches 0.84.

Limits: FOIA file contains only resolved loans (selection bias in recent vintages); no bureau, cash-flow or collateral data; pricing parameters illustrative. Portfolio project, not a production model.

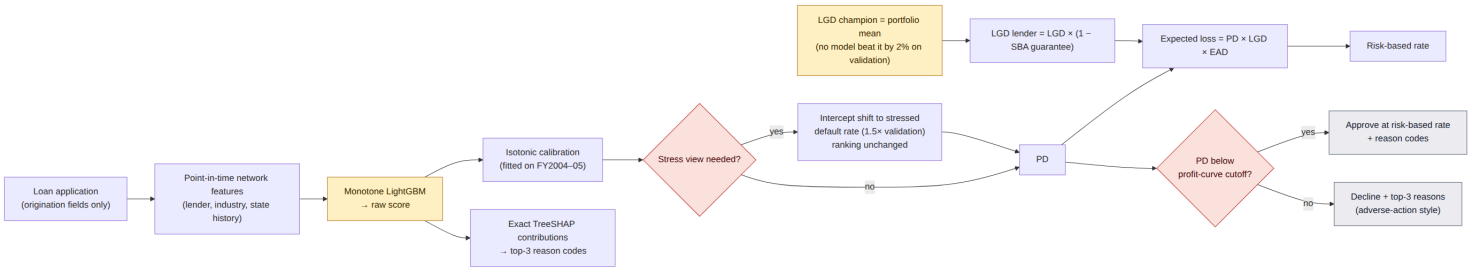
# smb-credit-pd-lgd · Pipeline: sources, steps and logic

Blue = external data source · purple = simulated data · lavender = code step (file named) · yellow = model · red = check or gate · grey = output.



# smb-credit-pd-lgd · Model and decision logic for one loan

Blue = external data source · purple = simulated data · lavender = code step (file named) · yellow = model · red = check or gate · grey = output.



# smb-credit-pd-lgd · Production design (proposed, not implemented)

Blue = external data source · purple = simulated data · lavender = code step (file named) · yellow = model · red = check or gate · grey = output.

