Chapter 5 - Empirical Results: High Tail Recall with Measured Sacrifice

Quantifying the economic cost of safety in Slow-21 tail gating

Abstract. This chapter reports empirical performance for the *Slow-21* tail gate, a survival-AFT model used as a high-precision classifier for identifying long-tail listings ("slow" meaning time-to-sale >= 504 hours). We introduce and operationalize the *sacrifice* family of metrics---$\mathbf{SAC_{LT10}}$ and $\mathbf{SAC_{MID}}$---that quantify the opportunity cost of safety: how often the gate incorrectly labels fast and mid-speed deals as tail and therefore suppresses them. On an out-of-time evaluation slice (SOLD last 7 days, N=941), the gate achieves **Precision**=0.8314, **Recall**=0.8614, **F1**=0.8462 for the slow class, while keeping $\mathbf{SAC_{LT10}}$=0.0059 (0.59%) and $\mathbf{SAC_{MID}}$=0.2688 (26.88%).

5.1 Experimental protocol and evaluation design

The Slow-21 gate is trained to support two downstream decisions that are economically coupled but statistically distinct: (i) **deal selection** (avoid stale inventory that ties up attention and capital), and (ii) **duration modeling** for the survival-AFT estimator that predicts time-to-sale. In both cases, the system must be *strictly* time-consistent: every feature used at inference must be computable at the listing's decision time *t0*.

The empirical protocol follows a strict temporal split: older SOLD examples (plus censored examples) constitute the training population; a recent SOLD window immediately preceding evaluation is reserved for calibration; and the most recent SOLD window is held out for evaluation. This layout ensures that hyperparameter selection, early stopping, and calibration are performed without looking into the future.

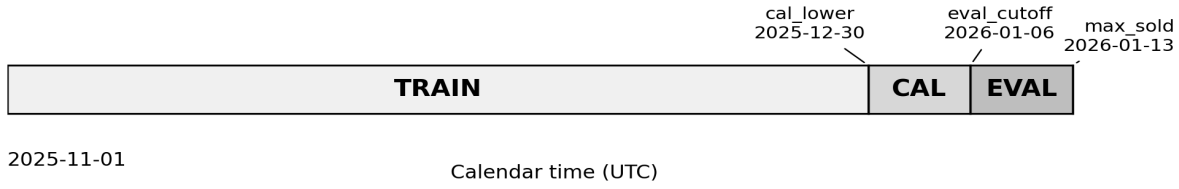

Figure 5.1 - Temporal protocol used for train / calibration / evaluation (UTC).

Label definition.

We define the Slow-21 classification target by thresholding the observed sale duration T (in hours) at tau=504h (21 days):

$$y_{\text{slow}} = \mathbf{1}\{T \geq \tau\}, \qquad \tau = 504\,\text{h}$$

Operationally, the gate is trained with an **Accelerated Failure Time** (AFT) objective and outputs a continuous prediction **T_hat** in hours. Classification is obtained by applying the same threshold tau to predictions: predict *slow* if T_hat >= tau. This approach is deliberate: the model preserves an ordering over time-to-sale that can be reused downstream, while still permitting a hard decision boundary for gating.

5.2 Metrics: accuracy, calibration, and the economics of sacrifice

Standard classification metrics (precision/recall/F1) for the slow class measure how effectively the gate separates long-tail "zombies" from non-tail inventory. However, marketplace decision quality depends on **who** is misclassified, not only **how many**. Misclassifying a listing that would have sold in 24-48 hours has a very different economic cost than misclassifying one that would have sold in 18 days.

Precision / recall / F1.

For completeness, the core metrics are computed on the slow (positive) class:

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad F_1 = \frac{2\,\mathrm{Prec}\,\mathrm{Rec}}{\mathrm{Prec}+\mathrm{Rec}}$$

Here TP is true slow predicted slow, FP is fast predicted slow, FN is slow predicted fast, and TN is true fast predicted fast.

Sacrifice metrics.

We define **sacrifice** as a false-positive slow prediction within a fast-moving speed band. Two bands are tracked: **LT10** (durations < 240h, i.e., < 10 days) and **MID** (240h <= duration < 504h, i.e., 10-21 days). These bands correspond to (a) highly desirable deals that are typically worth surfacing aggressively (LT10), and (b) intermediate-speed deals where a model can trade off more aggressively to protect tail recall (MID).

$$\mathrm{SAC}_{LT10} = \frac{1}{N_{LT10}} \sum_{i\,:\,T_i < 240} \mathbf{1}\{\hat{T}_i \geq \tau\}, \quad \mathrm{SAC}_{MID} = \frac{1}{N_{MID}} \sum_{i\,:\,240 \leq T_i < \tau} \mathbf{1}\{\hat{T}_i \geq \tau\}$$

Intuitively: $\mathrm{SAC}_{\mathrm{LT10}}$ answers "among listings that would sell in <10 days, what fraction does the gate incorrectly suppress as tail?". $\mathrm{SAC}_{\mathrm{MID}}$ answers the same question for the 10-21 day band. These metrics are *directional* and support explicit business constraints (for example, maintaining $\mathrm{SAC}_{\mathrm{LT10}}$ below a fixed cap) while maximizing tail recall.

5.3 Primary results

Table 5.1 summarizes performance on the out-of-time evaluation slice (SOLD last 7 days) and the calibration slice (recent SOLD used to fit the isotonic calibrator). Confidence intervals shown for precision and recall are Wilson 95% intervals.

Slice	N	Slow prev	Prec (Slow) (95% CI)	Rec (Slow) (95% CI)	F1 (Slow)	SAC_ LT10	SAC_ MID
Eval (SOLD last 7d)	941	17.64%	0.8314 [0.7683, 0.8800]	0.8614 [0.8007, 0.9059]	0.8462	0.0059	0.2688
Calibration (recent SOLD)	853	20.05%	0.6532 [0.5921, 0.7097]	0.9474 [0.9030, 0.9721]	0.7733	0.0143	0.6341

Table 5.1 - Slow-21 gate metrics (slow class = duration >= 504h).

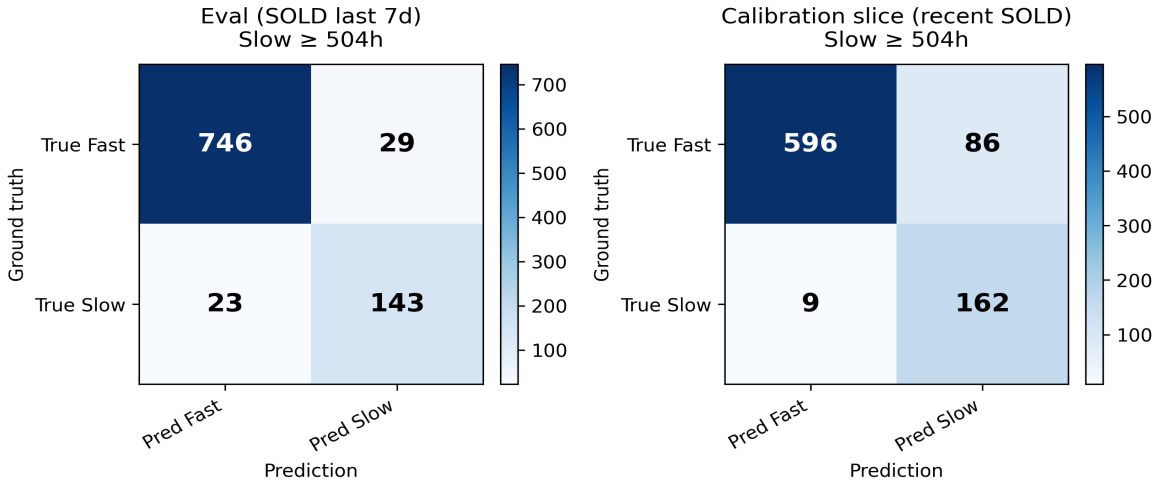

Figure 5.2 - Confusion matrices for evaluation and calibration slices (slow >= 504h).

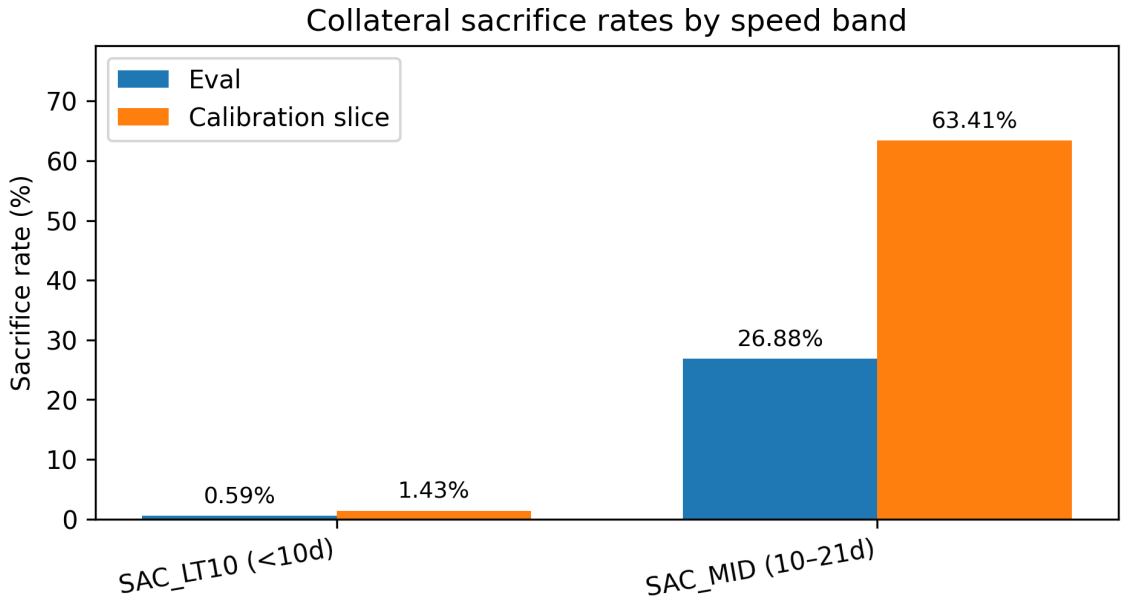

Figure 5.3 - SAC metrics quantify collateral suppression by speed band.

Interpretation.

On the evaluation slice (N=941), slow prevalence is 17.64%. The gate captures the majority of slow listings (Recall=0.8614) while remaining selective (Precision=0.8314). Critically, SAC_{LT10} is only 0.0059 (approx. 0.59%), indicating that the model almost never suppresses listings that would sell within 10 days. The model instead concentrates its sacrifice in the MID band (SAC_{MID}=0.2688), which is the intended trade-off region when maximizing long-tail recall.

The calibration slice reports higher recall (0.9474) but substantially higher sacrifice (SAC_MID=0.6341). Because the calibrator is fit on this slice, these numbers should be interpreted as a stress test of how aggressive a tail-favoring calibration can become when optimized for boundary fidelity. The out-of-time evaluation slice remains the primary indicator of generalization.

5.4 Optimization objective and best hyperparameters

The gate is optimized via Bayesian hyperparameter search (Optuna) using a custom evaluation metric aligned with the business objective: **F1 on the slow class at tau=504h**. The underlying learner is an XGBoost AFT model, which supports censoring and provides a continuous time prediction T_hat. During search, sample weights emphasize tail examples (slow_tail_weight and very_slow_tail_weight), and a boundary-aware calibration weighting scheme increases fidelity near the decision threshold.

Hyperparameter	Value
learning_rate	0.0123259
num_leaves	22
min_data_in_leaf	27
feature_fraction	0.574384
bagging_fraction	0.949383
lambda_l2	9.9161
aft_scale	1.47073
slow_tail_weight	6.30908
very_slow_tail_weight	6.34946
boundary_focus_k	1.14935
boundary_focus_sigma	5.42073

Table 5.2 - Best hyperparameters from Optuna (slow-class F1 objective).

In this run, the best slow-class F1 on the evaluation objective was **0.8462**. Notably, the fitted tail weights are substantial (about 6x), reflecting the strong class imbalance and the need to prioritize long-tail identification over minimizing error on the dominant fast class.

5.5 What drives decisions: feature attribution at evaluation time

The top gain-importance features provide a sanity check on the learned decision policy. The two strongest signals---ptv_anchor_strict_t0 and ptv_anchor_smart---account for more than half of the total gain, indicating that the model heavily leverages leak-proof anchored price-to-value structure at t0. This is consistent with the hypothesis that mispricing relative to cohort priors is a dominant driver of long-tail risk.

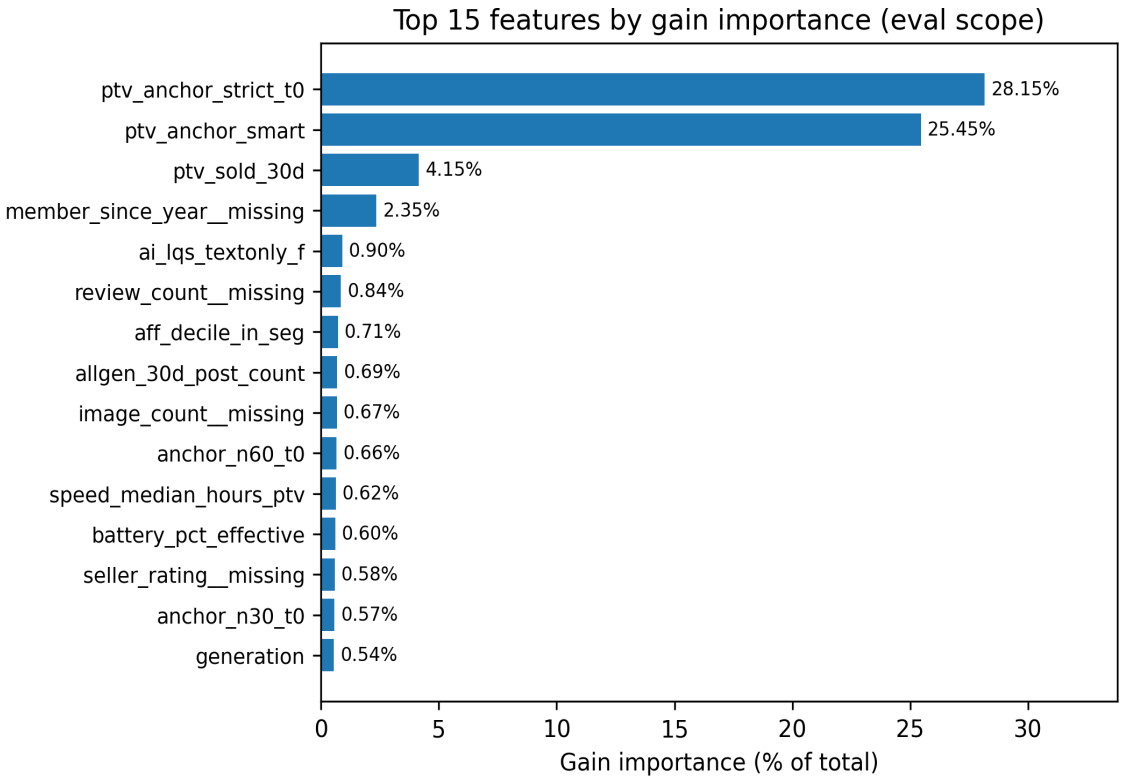

Figure 5.4 - Top features by gain importance (evaluation scope).

Secondary drivers include seller trust indicators (missingness flags), device condition proxies (battery and damage), and coarse supply-side dynamics (post counts). These features are precisely the categories expected to influence time-to-sale in a frictional peer-to-peer marketplace, and their presence supports the construct validity of the system.

5.6 Economic interpretation: why SAC matters

A tail gate is not merely a classifier; it is an **allocation policy**. Predicting slow suppresses or deprioritizes a listing and therefore has an opportunity cost. SAC metrics convert this cost into a measurable quantity by conditioning on realized speed bands. In particular, SAC_LT10 measures the rate at which the system discards the very inventory that creates the highest user satisfaction and the strongest marketplace conversion.

The empirical result SAC_LT10=0.0059 on the evaluation slice indicates that the system can increase tail recall without materially harming the fast-deal funnel. This is the key property that makes the gate deployable: tail capture is achieved primarily through trading off the mid band, not by discarding the best deals.

5.7 Recommended next experiments

The results support deployment-grade tail capture, but the sacrifice framework is intentionally designed to enable systematic experimentation. The next steps that most directly strengthen causal confidence are:

• Threshold sweep: report (Recall_slow, Precision_slow, SAC_LT10, SAC_MID) as a function of tau and/or the decision threshold applied to T_hat.

• Ablation: re-train with anchors removed, then with vision removed, to quantify marginal contribution under SAC constraints.

• Segment stability: compute metrics by (generation x super_metro x storage bucket) to ensure SAC_LT10 remains low in all high-traffic segments.

• Calibration audit: evaluate calibrator generalization on a second held-out window to rule out overfitting to the calibration slice.

Finally, because the downstream survival model consumes gate outputs, the most meaningful end-to-end evaluation is economic: measure improvement in realized ROI / conversion under a policy that deprioritizes predicted-slow listings while keeping SAC_LT10 bounded. This chapter establishes the measurement foundation for that experiment.

5.8 Summary

The Slow-21 gate achieves high long-tail recall on an out-of-time slice while maintaining exceptionally low collateral suppression of sub-10-day deals. The sacrifice metrics provide a principled bridge between statistical model selection and marketplace economics, enabling transparent, constraint-aware optimization as the system evolves.