CRAFT-GBM: SELECTING BEST FEATURES FOR GRADIENT-BOOSTED MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature subsampling is widely used to accelerate gradient-boosted trees, yet standard approaches choose the number of features in advance and largely ignore the split-gain evidence accumulated during training. We introduce CRAFT-GBM (Confidence-Ranked Adaptive Feature Tracking), a native LightGBM extension that uses this evidence to adapt which—and, by default, how many—features are searched in each boosting round. CRAFT-GBM transforms root-split gains into bounded observations and constructs joint time-uniform bounds that remain valid under adaptive, non-stationary training histories. In its default endogenous mode, after initialization, CRAFT-GBM retains exactly the features that cannot yet be ruled out as best, allowing the selected feature-set size K_t to emerge from the observed training history rather than from a pre-specified fraction. With probability at least $1 - \delta$, every feature with the largest expected gain is retained jointly across all rounds after initialization. When a feature fraction f is specified instead, CRAFT-GBM ranks features by their interval point estimates and provides an approximate top-K guarantee for the corresponding feature budget K_f. We implement CRAFT-GBM directly within LightGBM while preserving its native feature-bundling representation, and evaluate it across 14 public and synthetic dataset families together with four noise-augmented settings. In the fully endogenous setting, CRAFT-GBM reduces per-fit training time by 39.74% while achieving 1.04% lower test loss in the geometric mean across all 18 RQ1 experiments. Under pre-specified feature fraction f, Exogenous CRAFT-GBM also yields aggregate training-time reductions up to 13.52 percentage points larger than random feature subsampling, with no increase in aggregate test loss relative to full-feature LightGBM. These results show that accumulated training evidence can support adaptive feature search with statistical guarantees while retaining the practical advantages of native gradient boosting.

1 INTRODUCTION

Gradient-boosted decision trees (GBDTs) remain a strong choice for tabular learning (Grinsztajn et al., 2022; McElfresh et al., 2023), but high-dimensional training is expensive across hyperparameter search, cross-validation and retraining. Histogram-based boosting builds and scans every searched feature's split statistics each round (Ke et al., 2017), although many features are weak, redundant, or rarely competitive for the best split. Random feature subsampling reduces this work, but fixes the fraction in advance and ignores earlier split-gain evidence. Related work addresses ensemble-wide sparsity (Ibrahim et al., 2024) and feature-rank uncertainty (Neuhof & Benjamini, 2024). We ask: *can a boosting algorithm use its accumulated split-gain evidence to decide which features still need to be searched and, by default, how many?*

Historical gain rankings alone are an obvious but insufficient solution. A valid screening rule must address four challenges: (C1) native split gains are unbounded and their scale drifts as the ensemble fits; (C2) a feature's gain is observed only when it is searched, so missing observations are caused by earlier screening decisions; (C3) gains are non-stationary because they depend on the current learner; and (C4) screening decisions and the stopping round depend on the training history, so guarantees must hold at every round and for all features at once. Classical best-arm and subset identification handles adaptive sampling and stopping under fixed arm means (Even-Dar et al., 2006; Kalyanakrishnan et al., 2012). Racing handles adaptive sampling (Tiwari et al., 2022), while rising-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bandit identification assumes nondecreasing, concave rewards (Mussi et al., 2024). Our target needs no such shape assumptions.

We introduce CRAFT-GBM (Confidence-Ranked Adaptive Feature Tracking), a native LightGBM extension that treats per-round feature screening as sequential inference. Its design is theory-driven: each component is the construction a formal guarantee requires under one of these challenges. For (C1), a feature gain transformation maps each searched feature's best root-split gain to $[0, 1]$ using a scale fixed before the round, so observations are bounded and comparable across rounds. For (C2) and (C3), each feature's *accumulated expected gain* averages its conditional expected transformed gains over the rounds in which it was actually searched, a target that follows the evolving learner without imputing gains for skipped rounds. For (C4), empirical-Bernstein confidence sequences (Howard et al., 2020; 2021) bound these targets jointly over all features and rounds under history-dependent selection and stopping, without independence or stationarity assumptions (Theorem 1). The empirical-Bernstein construction uses squared prediction errors to control deviations from conditional expected feature gains, with confidence interval widths adapting to both native LightGBM's GOSS sampling variance and the accuracy of the historical predictor.

In CRAFT-GBM, two screening rules turn these bounds into a search policy. In the default *Endogenous* mode, after a 20-round full-feature warm-up, CRAFT-GBM searches exactly the features that the joint intervals still allow to be best (Proposition 1). The feature budget K_t is therefore set by the evidence: a feature stays in while it is uncertain, is skipped once it is confidently dominated, and can re-enter later. With probability at least $1 - \delta$, jointly over all rounds, no feature with the largest expected gain is dropped and every searched feature has a bounded gap to the best (Theorem 2); the set-level guarantee holds for the random budget K_t without an additional union bound (Corollary 1). When the user specifies a feature fraction f, the *Exogenous* mode searches the corresponding K_f features with the largest point estimates and bounds their gap to the best size-K set (Theorem 3).

CRAFT-GBM runs inside LightGBM's training loop and uses its Exclusive Feature Bundling layout unchanged. Both modes are compared with LightGBM-Fixed's random subsampling and the native samplers of XGBoost and CatBoost, using LightGBM-Full (no feature selection) as the baseline comparison. RQ2 additionally includes two ablations: greedy mean-gain ranking (evidence without uncertainty) and exposure balancing (exploration without evidence). Across all 18 RQ1 experiments, Endogenous CRAFT-GBM reduces aggregate training time by 39.7% relative to the baseline LightGBM-Full, and its aggregate test loss is 1.0% lower. At feature fraction $f = 0.10$, Exogenous CRAFT-GBM reduces aggregate training time by 62.7% with essentially unchanged test loss (-0.04%). The two ablations are about as fast, reducing training time by 63.8% and 64.7%, but they raise aggregate test loss by 12.7% (greedy mean-gain ranking) and 5.7% (exposure balancing). Random subsampling by LightGBM-Fixed reduces aggregate training time by 30.9% in RQ1 and 49.2% at $f = 0.10$ in RQ2. All values in Figures 1 and 2 are relative to the baseline LightGBM-Full; test loss is log loss for classification and RMSE for regression.

Our contributions are:

1. **Sequential inference for adaptive feature screening.** We formulate feature screening in boosting as inference on adaptively observed, path-dependent gains, and introduce a predictable bounded-gain construction and training-path target that make joint time-uniform confidence sequences applicable (Theorem 1).
2. **Evidence-determined feature budgets with guarantees.** The Endogenous rule lets uncertainty determine how many features are searched and provides simultaneous best-feature retention and near-best guarantees (Theorem 2). For a user-specified budget, the Exogenous rule provides an approximate top-K guarantee (Theorem 3).
3. **Native integration and empirical efficiency.** We integrate CRAFT-GBM directly into LightGBM while preserving its native EFB representation. Across all 18 RQ1 experiments, it reduces aggregate training time by 39.7% with 1.0% lower aggregate test loss relative to full-feature LightGBM.

2 METHOD

Setup and notation. Let $p \geq 1$ count the valid features after native preprocessing by LightGBM, indexed by $[p] = \{1, \ldots, p\}$. At boosting round t, CRAFT-GBM selects a set of features (columns)

$S_t \subseteq [p]$, with $K_t = |S_t|$, before the native GOSS sampling of rows. Each evaluated feature contributes its best eligible root-split gain $G_{j,m,t}$, whether or not it wins the split. The observed root-split gains $G_{j,m,t}$ are mapped to the bounded interval $[0, 1]$ via Feature Gain Transformation (FGT). A feature's *expected gain* $\overline{\nu}_{j,t-1}$ is the conditional expectation of gain averaged across boosting rounds in which the feature is selected, whereas its *empirical mean* $\widehat{\mu}_{j,t-1}$ is the average of its observed values. A best feature has the *largest expected gain* $\nu_t^\star$, not necessarily the largest observed gain in the current round. At round t, confidence sequences provide confidence intervals $[L_{j,t}, U_{j,t}]$ of expected gains $\overline{\nu}_{j,t-1}$ jointly across features and rounds through round $t-1$, which endogenous feature selection uses to retain a set of possible-best features S_t, whose size K_t is a random variable. When a feature fraction f is specified, Exogenous CRAFT-GBM selects the $K_f = \max\{1, \lfloor fp + 1/2 \rfloor\}$ features with the largest *interval point estimates* $c_{j,t}$. Algorithm 1 in Appendix F gives the selection and update procedure.

2.1 Feature Gain Transformation (FGT)

Let $M \geq 1$ be the fixed number of native output trees in one boosting round. For selected feature j, output component m, and round t, define

$$G_{j,m,t} = \max\left(\{0\} \cup \left\{\Delta_{m,t}^{\text{root}}(j, \theta) : \theta \in \Theta_{j,m,t}^{\text{root}}\right\}\right). \tag{1}$$

Here $\Theta_{j,m,t}^{\text{root}}$ is the native eligible candidate set $\Theta_{j,m,t}^{\text{root}}$ for feature j, and Δ is LightGBM's net eligible split gain $\Delta_{m,t}^{\text{root}}(j, \theta)$ calculated by its native tree algorithm (Chen & Guestrin, 2016; Ke et al., 2017). The union with zero covers a selected feature with no positive eligible candidate. Hence, given a feature, Equation (1) gives the feature's best root opportunity $G_{j,m,t}$ in a boosting round t, evaluated at the current learner state, row sample, and native candidate eligibility criteria without requiring that feature to win against other features in that round.

For output component m, initialize $s_{m,1} = \perp$. Thereafter, let the scale $s_{m,t} \in (0, \infty) \cup \{\perp\}$ be fixed before round t, and set

$$\phi(g; s) = \begin{cases} g/(g+s), & s > 0, \\ \mathbf{1}\{g > 0\}, & s = \perp, \end{cases} \qquad R_{j,t} = \frac{1}{M} \sum_{m=1}^{M} \phi(G_{j,m,t}; s_{m,t}). \tag{2}$$

Then $R_{j,t} \in [0, 1]$. For $s > 0$, the transformation $\phi(g; s)$ is strictly increasing, and $\phi(ag; as) = \phi(g; s)$ for every $a > 0$. After an accepted round, the next scale $s_{m,t+1}$ for output m is the median of

$$\{G_{j,m,t} : j \in S_t,\ G_{j,m,t} > 0\},$$

using the midpoint for an even number of values. This multiset contributes at most one value per selected feature. If it is empty, the previous scale $s_{m,t}$ is retained. Current gains cannot affect their same-round transformation $\phi(G_{j,m,t}; s_{m,t})$.

2.2 Uniform Moment Statistics (UMS)

Uniform Moment Statistics (UMS) maintain an equal-weight transformed-gain sum $A_{j,t}$, an observation count $N_{j,t}$, and the squared-error sum $V_{j,t}$ used below for each feature over its selected rounds. Let

$$I_{j,t} = \mathbf{1}\{j \in S_t\}, \qquad Y_{j,t} = \begin{cases} R_{j,t}, & j \in S_t, \\ 0, & j \notin S_t. \end{cases}$$

When a feature is omitted from boosting round t, the assigned zero is for bookkeeping, not to indicate that the counterfactual gain is zero. Then, starting from $A_{j,1} = N_{j,1} = 0$, for $t \geq 2$ define

$$A_{j,t} = A_{j,t-1} + Y_{j,t}, \qquad N_{j,t} = N_{j,t-1} + I_{j,t}. \tag{3}$$

Equivalently,

$$A_{j,t} = \sum_{s=2}^{t} Y_{j,s}, \qquad N_{j,t} = \sum_{s=2}^{t} I_{j,s}.$$

For $N_{j,t} > 0$, define the empirical mean $\widehat{\mu}_{j,t}$ over feature j's selected rounds as

$$\widehat{\mu}_{j,t} = \frac{A_{j,t}}{N_{j,t}}.$$

Let $(\mathcal{F}_t)_{t\geq 1}$ describe the *training history* $\mathcal{F}_t$ through each accepted round. $\mathcal{F}_{t-1}$ contains the fixed data and configuration, the fixed feature mapping and native EFB representation, every committed learner and CRAFT-GBM update through $t-1$, the scale $s_{m,t}$ for round t, and all earlier GOSS draws before accepted round t. Selection is completed before the current-round GOSS sampling. Hence, $I_{j,t}$ is $\mathcal{F}_{t-1}$-measurable. The recorded value $Y_{j,t}$ is $\mathcal{F}_t$-measurable. In particular, $\mathcal{F}_1$ contains the data, configuration, feature map, and native EFB representation fixed before training, together with the committed round-1 learner F_1 and the scales $s_{m,2}$ for round 2. It describes the history $\mathcal{F}_1$ after round 1 and before round 2.

Define the conditional gain $\nu_{j,t}$ and, for $N_{j,t} > 0$, its accumulated average $\overline{\nu}_{j,t}$:

$$\nu_{j,t} = \mathbb{E}[Y_{j,t} \mid \mathcal{F}_{t-1}], \qquad \overline{\nu}_{j,t} = \frac{\sum_{s=2}^{t} \nu_{j,s}}{N_{j,t}}. \tag{4}$$

We call $\overline{\nu}_{j,t}$ the *expected gain* $\overline{\nu}_{j,t}$, while $\widehat{\mu}_{j,t}$ is its empirical estimator $\widehat{\mu}_{j,t}$. The target $\overline{\nu}_{j,t}$ is the expected transformed root gain $\overline{\nu}_{j,t}$, conditional on the preceding training history $\mathcal{F}_{s-1}$, averaged over the rounds in which that feature was evaluated. If $I_{j,s} = 0$, then $Y_{j,s} = 0$ and consequently $\nu_{j,s} = 0$, so the target averages conditional transformed root gains $\nu_{j,s}$ only over that feature's realized selected rounds. Different features may therefore be evaluated in different learner states. Predictability and $0 \leq Y_{j,t} \leq I_{j,t}$ imply $0 \leq \overline{\nu}_{j,t} \leq 1$. This target need not be stationary: it is an average of the conditional expectations over the accumulated rounds in which the feature was selected, not an instantaneous or recent-gain target.

2.3 Joint Time-Uniform Bounds

The confidence bounds $[L_{j,t}, U_{j,t}]$ adapt to how well past gains predict subsequent gains. Before observing round t, form the smoothed historical prediction $a_{j,t} = (1/2 + A_{j,t-1})/(1 + N_{j,t-1})$. Accumulate its squared errors:

$$V_{j,u} = \sum_{t=2}^{u} (Y_{j,t} - I_{j,t} a_{j,t})^2. \tag{5}$$

Only selected features contribute, since $Y_{j,t} = I_{j,t} = 0$ for an omitted feature. Crucially, $a_{j,t}$ is computed *before* the current gain updates A and N. This error sum $V_{j,u}$ is observable and does not assume stationary gains or estimate an unobserved gain for an omitted feature.

Set the per-tail error level $\alpha = \delta/(2p)$. Freeze a positive mixture scale ρ before round 1 using a prespecified reference count n_{ref} (default $n_{\mathrm{ref}} = 300$), independently of the stopping rule or any maximum-round ceiling (Appendix B). Let $u_\rho(v; \alpha)$ be the gamma-exponential mixture boundary $u_\rho(v; \alpha)$ defined in that appendix. The radius $r_{j,u}$ is

$$r_{j,u} = \frac{u_\rho(V_{j,u}; \delta/(2p))}{N_{j,u}}, \qquad N_{j,u} > 0. \tag{6}$$

This is an empirical-Bernstein confidence sequence (Howard et al., 2021, Theorem 4 and Proposition 9): smaller accumulated prediction errors $V_{j,u}$ give smaller radii $r_{j,u}$ at the same observation count $N_{j,u}$. It does not require replacing a worst-case variance by a sample variance in an otherwise unchanged formula.

Theorem 1 (Joint time-uniform bounds)**.** *Condition on $\mathcal{F}_1$, including the native feature map, native EFB representation, realized finite p, and the positive ρ frozen before round 1. For every $t \geq 2$, suppose $I_{j,t}$ is $\mathcal{F}_{t-1}$-measurable, $Y_{j,t}$ is $\mathcal{F}_t$-measurable, and $0 \leq Y_{j,t} \leq I_{j,t}$. Use the predictable predictions $a_{j,t}$ and error sums $V_{j,u}$ in Equation* (5)*, and the radii $r_{j,u}$ in Equation* (6)*. Almost surely,*

$$\Pr(\forall j \leq p,\ \forall u \geq 2 : N_{j,u} > 0 \Longrightarrow |\widehat{\mu}_{j,u} - \overline{\nu}_{j,u}| \leq r_{j,u} \mid \mathcal{F}_1) \geq 1 - \delta.$$

For each feature, two nonnegative mixture supermartingales control the upper and lower errors over all rounds through Ville's inequality. A union bound over these $2p$ events explains $\alpha = \delta/(2p)$, and fixing the mixture scale ρ consumes no additional error probability. The theorem therefore covers every boosting round, any history-adapted stopping endpoint, and valid continuation beyond n_{ref}, on the condition that the fixed feature universe $[p]$, bounded transformation, and other stated invariants persist. No independence is required among features, native EFB bundles, GOSS sampling draws, or multiclass output trees.

2.4 ENDOGENOUS FEATURE SELECTION

Endogenous feature selection can begin at any boosting round $t^\star \geq 3$. Preceding rounds $1 \leq t < t^\star$ use all features $S_t = [p]$: round 1 initializes the transformation scales $s_{m,2}$, and rounds $2 \leq t < t^\star$ accumulate confidence statistics $(A_{j,t}, N_{j,t}, V_{j,t})$. The theorems below hold for any such starting round $t^\star \geq 3$, and the Endogenous CRAFT-GBM uses $t^\star = 21$ by default.

At the beginning of round t, use only statistics through $t-1$. Define the bounded interval $[L_{j,t}, U_{j,t}]$

$$(L_{j,t}, U_{j,t}) = \begin{cases} \big(\max\{0, \widehat{\mu}_{j,t-1} - r_{j,t-1}\}, \min\{1, \widehat{\mu}_{j,t-1} + r_{j,t-1}\}\big), & N_{j,t-1} > 0, \\ (0,1), & N_{j,t-1} = 0. \end{cases} \tag{7}$$

Intersecting the raw interval $[\widehat{\mu}_{j,t-1} - r_{j,t-1}, \widehat{\mu}_{j,t-1} + r_{j,t-1}]$ with the known range $[0,1]$ preserves coverage. An unseen feature receives the complete parameter range. For endogenous rounds $t \geq 21$, let

$$C_t = \max_{1 \leq c \leq p} L_{c,t}, \qquad S_t = \{j : 1 \leq j \leq p,\ U_{j,t} \geq C_t\}, \qquad K_t = |S_t|. \tag{8}$$

All features are reconsidered each round, with omitted features retaining their statistics and confidence intervals unchanged to the next round, so re-entry is possible but not guaranteed. At least one feature attaining C_t is retained, so $1 \leq K_t \leq p$. All features, and hence every best feature, are retained during initialization. Let the joint rectangular confidence region $\mathcal{C}_t$ be

$$\mathcal{C}_t = \prod_{j=1}^{p} [L_{j,t}, U_{j,t}].$$

Proposition 1 (S_t is the set of possible-best features in round t)**.** *For $t \geq 21$ and any nonempty intervals in $\mathcal{C}_t$,*

$$S_t = \left\{ j : \exists v \in \mathcal{C}_t \text{ such that } v_j = \max_{1 \leq c \leq p} v_c \right\}.$$

Thus S_t is the smallest set containing every feature that can be a maximizer for some parameter vector allowed by the rectangular region $\mathcal{C}_t$.

For $t \geq 21$, define

$$\nu_t^\star = \max_{j \in [p]} \overline{\nu}_{j,t-1}, \qquad \mathcal{M}_t^\star = \arg\max_{j \in [p]} \overline{\nu}_{j,t-1}, \qquad r_t^\star = \min_{j \in \mathcal{M}_t^\star} r_{j,t-1}.$$

Theorem 2 (Endogenous feature selection guarantee)**.** *Conditional on $\mathcal{F}_1$, with probability at least $1-\delta$, the following statements hold jointly for every boosting round $t \geq 21$:*

1. *No true maximizer is omitted:*

$$\mathcal{M}_t^\star \subseteq S_t, \qquad K_t \geq |\mathcal{M}_t^\star|.$$

2. *For every omitted feature $j \notin S_t$, there is a retained feature $c \in S_t$ such that*

$$\overline{\nu}_{c,t-1} > \overline{\nu}_{j,t-1}.$$

3. *Every retained feature $j \in S_t$ is approximately best:*

$$0 \leq \nu_t^\star - \overline{\nu}_{j,t-1} \leq 2r_{j,t-1} + 2r_t^\star. \tag{9}$$

Consequently,

$$\begin{aligned} 0 &\leq \nu_t^\star - \frac{1}{K_t} \sum_{j \in S_t} \overline{\nu}_{j,t-1} \\ &\leq 2r_t^\star + \frac{2}{K_t} \sum_{j \in S_t} r_{j,t-1} \leq 4 \max_{j \in S_t} r_{j,t-1}. \end{aligned} \tag{10}$$

4. *If $K_t = 1$, the sole retained feature is the unique maximizer. More generally, if $\mathcal{M}_t^\star = \{1, \ldots, p\}$, exact recovery is automatic. Otherwise the sufficient separation condition*

$$\max_{j \notin \mathcal{M}_t^\star} U_{j,t} < \max_{c \in \mathcal{M}_t^\star} L_{c,t} \tag{11}$$

implies $S_t = \mathcal{M}_t^\star$.

Hence, the theorem guarantees selection of all features "not ruled out as best" in each boosting round.

Corollary 1 (Guarantee of selecting a set of possible-best features S_t)**.** *At every round covered by Theorem 2, let*

$$S_t^{\star,K_t} \in \underset{S\subseteq\{1,\dots,p\},\ |S|=K_t}{\arg\max} \sum_{j\in S} \overline{\nu}_{j,t-1}.$$

Then, jointly,

$$0 \le \frac{1}{K_t}\left(\sum_{j\in S_t^{\star,K_t}} \overline{\nu}_{j,t-1} - \sum_{j\in S_t} \overline{\nu}_{j,t-1}\right) \le 2\max_{j\in S_t} r_{j,t-1}. \tag{12}$$

No additional union bound over possible realized values of K_t is needed.

2.5 Exogenous Feature Selection

Users can instead fix a feature fraction f before training. After initialization, Exogenous CRAFT-GBM selects an approximate best-K_f set in each round, with K_f defined below.

Exogenous K through `feature_fraction`. A value $f = \texttt{feature_fraction} \in (0,1)$, fixed before training, instead specifies

$$K_f = \max\left\{1, \left\lfloor fp + \frac{1}{2}\right\rfloor\right\}, \tag{13}$$

using LightGBM's native feature-count rounding rule. Rounds 1 and 2 still evaluate all p features: round 1 initializes the FGT scales $s_{m,2}$, and round 2 supplies the first confidence observation $R_{j,2}$ for every feature. For every round $t \ge 3$, define the interval point estimate $c_{j,t}$ and half-width $h_{j,t}$:

$$c_{j,t} = \frac{L_{j,t} + U_{j,t}}{2}, \qquad h_{j,t} = \frac{U_{j,t} - L_{j,t}}{2}.$$

CRAFT-GBM selects exactly the K_f features with the largest $c_{j,t}$. This confidence interval point estimate $c_{j,t}$ is generally different from the empirical mean $\widehat{\mu}_{j,t-1}$ because bounding the endpoints $L_{j,t}, U_{j,t}$ can make the interval $[L_{j,t}, U_{j,t}]$ asymmetric around that mean $\widehat{\mu}_{j,t-1}$. Equal interval point estimates $c_{j,t}$ are resolved by smaller observation count $N_{j,t-1}$, followed by smaller canonical feature identifier $\mathrm{id}(j)$. This tie rule favors a less-observed feature but does not alter the theorem below. Exogenous CRAFT-GBM feature selection is deterministic given the training history, so native LightGBM's parameter `feature_fraction_seed` has no effect, but the per-node parameter `feature_fraction_bynode` must remain one: additional node-level sampling could remove a selected feature at the root while its observation count $N_{j,t}$ was still incremented.

Let $S_t^{(f)}$ denote the resulting size-K_f set $S_t^{(f)}$ and let

$$S_t^{\star,K_f} \in \underset{S\subseteq\{1,\dots,p\},\ |S|=K_f}{\arg\max} \sum_{j\in S} \overline{\nu}_{j,t-1}.$$

Theorem 3 (Exogenous top-K selection)**.** *Under the assumptions of Theorem 1, conditional on $\mathcal{F}_1$, with probability at least $1-\delta$, jointly at every boosting round $t \ge 3$,*

$$\begin{aligned} 0 &\le \frac{1}{K_f}\left(\sum_{j\in S_t^{\star,K_f}} \overline{\nu}_{j,t-1} - \sum_{j\in S_t^{(f)}} \overline{\nu}_{j,t-1}\right) \\ &\le \frac{1}{K_f}\left(\sum_{j\in S_t^{\star,K_f}\setminus S_t^{(f)}} h_{j,t} + \sum_{j\in S_t^{(f)}\setminus S_t^{\star,K_f}} h_{j,t}\right) \le 2\max_{j\le p} h_{j,t} \le 1. \end{aligned} \tag{14}$$

For fixed f, this joint guarantee bounds the selected set's average expected-gain gap from the best size-K_f set $S_t^{\star,K_f}$ in that run, starting at round 3. Exogenous selection may omit a best feature even when at most K_f remain possible-best. The bound accounts for uncertainty in both selected and omitted features, while targets exclude unobserved omitted-round gains. No additional error allocation over rounds or sets is needed.

3 EXPERIMENTAL EVALUATION

- **RQ1:** Can Endogenous CRAFT-GBM reduce training time while maintaining test loss relative to competitor models, using the same pre-run-determined settings?
- **RQ2:** How does Exogenous CRAFT-GBM compare with competitor models across a set of pre-specified feature fractions, using LightGBM-Full as the baseline reference?

3.1 EXPERIMENTAL SETUP AND DESIGN

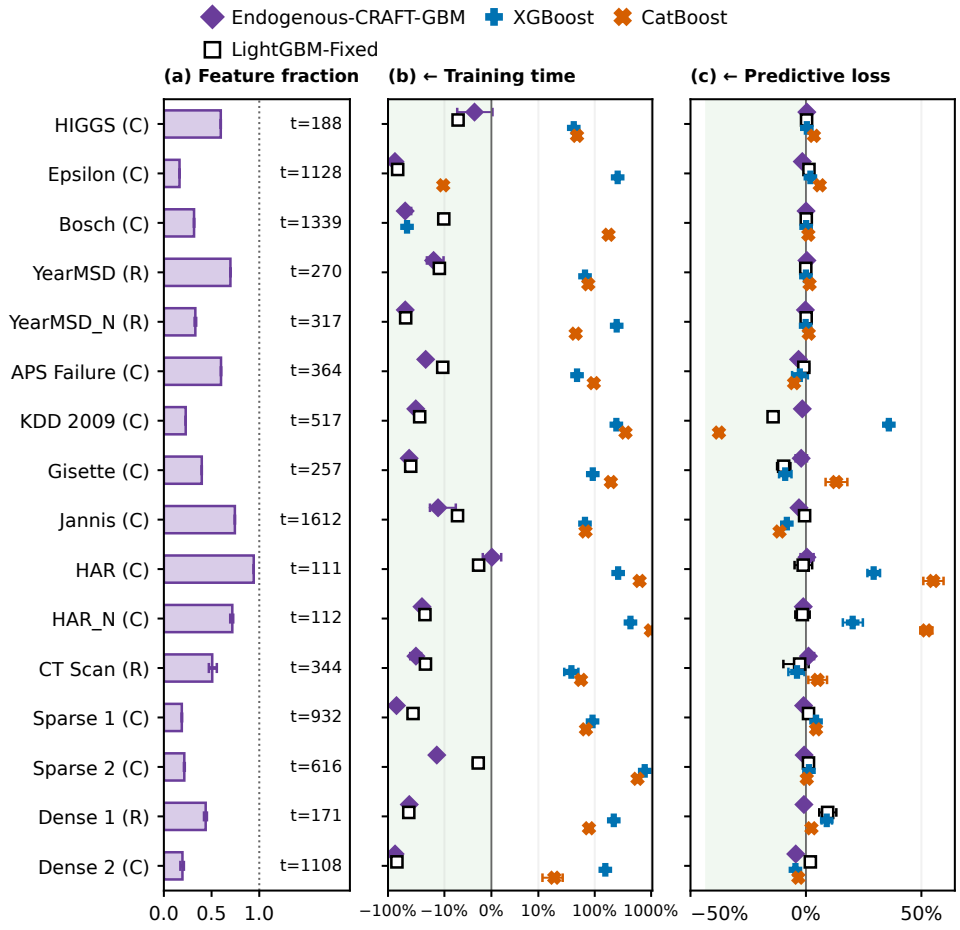
Datasets and evaluation. Ten public benchmark datasets cover binary/multiclass classification and regression. Four synthetic dataset families cover sparse inputs (Sparse 1), categorical indicators (Sparse 2), nonlinear regression (Dense 1), and redundant-feature classification (Dense 2), each with 2,000 features and 100,000 rows per training/test split. All competitor models share the same fixed training/test partitions for each dataset, with subject-disjoint HAR splits and patient-disjoint CT folds. Each experiment has five paired measurements. The test set is withheld from training and test labels are never used to select features or any other training settings. Test loss is log loss for classification (C) and RMSE for regression (R). RQ2 uses these 14 families. Figure 1 adds HAR_N and YearMSD_N, which are expanded HAR and YearMSD datasets with noise. This is done by appending eight row-permuted copies of every feature column, which expands the column size of each dataset by nine times. However, added row-permuted columns need not have zero gain and are expected to have augmented noise.

RQ1: endogenous feature selection. At the beginning of the training, for each dataset, one untimed Endogenous CRAFT-GBM pre-run is carried out to determine the total number of boosting rounds t and average feature fraction $\bar{f} = (pt)^{-1} \sum_{s=1}^{t} K_s$, which include initialization. Endogenous CRAFT-GBM is configured to stop independently at $\tau = \bar{t}$, which is after ten boosting rounds without a training-loss improvement exceeding $0.001L_0$ over its previous round's loss, where L_0 is initial training loss. The same pre-run-determined $\bar{t}$ and $\bar{f}$ are then used for all competitor models and the baseline LightGBM-Full model for fair comparison. LightGBM-Full, the baseline reference, uses all features for $\bar{t}$ boosting rounds. LightGBM-Fixed also runs for a total of $\bar{t}$ rounds, but randomly samples a fixed fraction $\bar{f}$ of features in each round. XGBoost (Chen & Guestrin, 2016) and CatBoost (Prokhorenkova et al., 2018) receive $\bar{f}, t$ through native samplers. Hence, both competitor models also match the feature and total running budgets of Endogenous CRAFT-GBM. All models use a learning rate of 0.05 without tuning. Figure 1 compares each competitor model's training time and test loss with those of LightGBM-Full.

RQ2: exogenous feature selection. Exogenous CRAFT-GBM, LightGBM-Fixed, XGBoost, CatBoost, and two ablations use $f \in \{0.10, 0.30, 0.50, 0.70, 0.90\}$. For each training dataset, a 20-trial Bayesian hyperparameter search tunes for LightGBM-Full's optimal rounds $T^\star$ in $[50, 2000]$ and optimal learning rate $\eta^\star$ in $[0.01, 0.20]$, minimizing loss on an 80/20 inner split. Then, all competitor models (Exogenous CRAFT-GBM, LightGBM-Fixed, XGBoost, CatBoost) are trained with the same hyperparameter-tuned $T^\star$ and $\eta^\star$ for each dataset, across all feature fractions $f \in \{0.10, 0.30, 0.50, 0.70, 0.90\}$. Their performance results in Figure 2 are all relative changes compared to the baseline LightGBM-Full. There are two ablation models running on the same settings. Mean ablation greedily ranks empirical gain means $\widehat{\mu}_{j,t-1}$ (evidence without uncertainty), whereas Exposure ablation selects the least-observed features in each boosting round (exploration without gain evidence).

Shared settings and reporting. All LightGBM-derived models share the same native LightGBM preprocessing, EFB bundling, and GOSS sampling at depth 6 and at most 64 leaves. CRAFT-GBM uses $\delta = 0.05$, with $n_{\text{ref}} = 300$ in RQ1 and $n_{\text{ref}} = 1000$ in RQ2. All models use equal CPU core counts and a standardized computational environment within each RQ. Total training time includes launch, loading, selection, boosting, and saving, but excludes preparation, pre-runs, tuning, and test prediction. In both Figure 1 and Figure 2, training time and test loss changes are reported as $100(r-1)\%$, where r averages **Competitor model / LightGBM-Full ratios** geometrically within and then equally across experiments. Hence, for all competitor models, **negative values indicate improvements in training time/test loss relative to LightGBM-Full**, and the opposite holds for positive values. LightGBM-Full is not plotted on both figures because it is used as the baseline comparison in the performance metric above (its ratios = 1 trivially).

3.2 RQ1: Endogenous Feature Selection

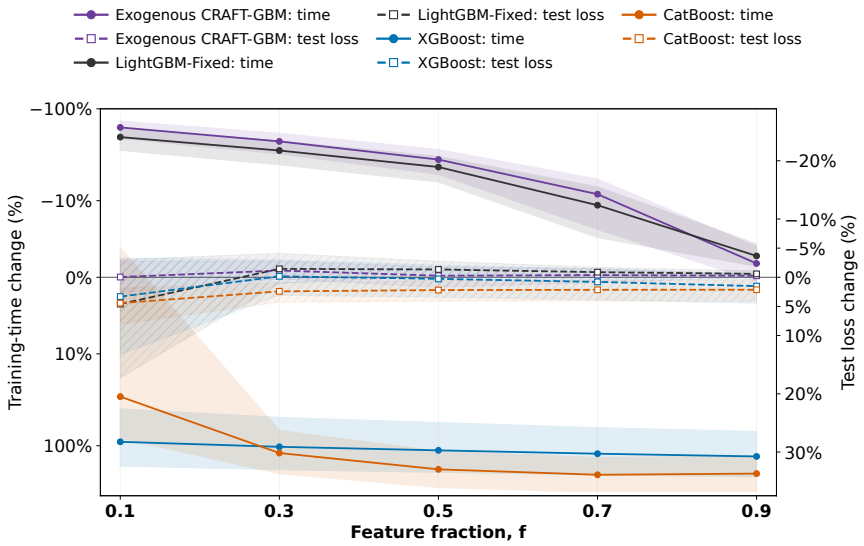


Model	16-experiment aggregate change		Wins across 16 experiments		
	Time (%)	Test loss (%)	Faster	Lower loss	Both
Endogenous CRAFT-GBM	-40.36	-1.10	15	11	11
LightGBM-Fixed	-30.65	-1.10	16	8	8
XGBoost	+135.77	+3.95	1	7	0
CatBoost	+140.52	+3.51	1	4	0

Figure 1: **Endogenous CRAFT-GBM reduces training time in 15 of the 16 displayed experiments.** Endogenous CRAFT-GBM reduces aggregate training time by 40.36% and test loss by 1.10%. LightGBM-Fixed saves 30.65% of training time and has the same aggregate loss change, so endogenous CRAFT-GBM delivers a further 9.70 percentage points of time reduction. Endogenous CRAFT-GBM improves both outcomes in $11/16$ experiments, compared with $8/16$ for random LightGBM-Fixed. Across almost all datasets, XGBoost and CatBoost are slower and incur higher test loss in some datasets.

Time savings span 3.6–75.4% in the 15 faster experiments, while HAR takes 0.11% longer. Loss falls 4.41% on Dense 2 but rises 1.09% on CT. Including both +4 expansions leaves the overall finding similar, with 39.74% less time and 1.04% lower loss across all 18 experiments. The HAR_N pre-run retains about 72% of features, so faster training need not remove every added column. Together, these results support Endogenous CRAFT-GBM as an evidence-determined alternative to fixing a feature fraction f in advance, without implying uniformly better predictions or recovery of only useful features.

3.3 RQ2: Performance Across Feature Fractions



Aggregate change across 14 source families (%)

f	**Exogenous CRAFT-GBM**		LightGBM Fixed		XGBoost		CatBoost		Mean ablation		Exposure ablation	
	Time	Loss	Time	Loss	Time	Loss	Time	Loss	Time	Loss	Time	Loss
0.10	-62.74	-0.04	-49.22	+4.60	+90.61	+3.31	+29.20	+4.44	-63.83	+12.69	-64.71	+5.71
0.30	-44.20	-1.16	-35.17	-1.44	+102.74	-0.17	+120.07	+2.42	-45.08	+7.73	-47.00	+2.16
0.50	-28.02	-0.27	-23.25	-1.34	+112.23	+0.28	+180.62	+2.20	-29.72	+4.40	-31.73	+1.94
0.70	-11.77	-0.38	-9.41	-0.88	+122.12	+0.79	+207.04	+2.15	-13.19	+3.53	-13.82	-0.99
0.90	-1.81	-0.26	-2.79	-0.59	+130.74	+1.52	+201.02	+2.14	-3.63	+0.41	-4.38	-0.41

Figure 2: **At $f = 0.10$, Exogenous CRAFT-GBM substantially reduces training time with essentially unchanged aggregate test loss.** Shaded bands show pointwise 95% dataset-resampling intervals, reflecting sensitivity to the composition of the 14 dataset families. At $f = 0.10$, Exogenous CRAFT-GBM reduces aggregate training time by 62.74% with essentially unchanged test loss (-0.04%). Random subsampling by LightGBM-Fixed saves 49.22% time but raises loss by 4.60%. Thus CRAFT-GBM delivers 13.52 percentage points more time reduction without the aggregate loss increase seen with LightGBM-Fixed. Greedy mean-gain ranking (Mean ablation) and exposure balancing of features (Exposure ablation) save the same training time (63.83% and 64.71%), but incur greater test loss especially at smaller feature fractions (12.69% and 5.71% more test loss at $f = 0.10$). This contrast is consistent with the expected benefit of combining gain evidence with uncertainty in CRAFT-GBM, especially when the feature budget K_f is restrictive.

Exogenous CRAFT-GBM also has the largest training-time reduction among non-ablated competitor models—LightGBM-Fixed, XGBoost, and CatBoost—dominating at $f \leq 0.70$, while its negative test loss changes indicate lower test loss than the baseline LightGBM-Full's across all feature fractions. As expected, the advantage gained by all competitor models narrows as feature fraction increases.

4 Conclusion

CRAFT-GBM turns accumulated split-gain evidence into statistically guided feature search, deciding which features to search and, by default, how many. Joint confidence bounds support possible-best retention and approximate top-K selection. Relative to LightGBM-Full, Endogenous CRAFT-GBM reduces aggregate training time by 39.74% with 1.04% lower test loss across all 18 RQ1 experiments. Exogenous CRAFT-GBM saves up to 62.74% of training time at $f = 0.10$ with essentially unchanged test loss, demonstrating efficient feature search even under restrictive feature budgets.

Ethics Statement

CRAFT-GBM is evaluated on public benchmarks and controlled synthetic data. Reuse of the public datasets is subject to their respective terms and sampling limitations. Its statistical guarantees concern feature selection during training and do not establish fairness or safety in downstream applications.

AI Use Statement

In this work, we used generative AI tools to generate synthetic data sets, assist in the writing of proofs, provide feedback on research methodology or experiments, clean and reformat datasets, and support qualitative and thematic data analysis. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, implement methods, or interpret results. Assistance with translation is not applicable to this work. We have reviewed all AI-assisted work, for example, by manually checking and approving all LLM-generated suggestions and feedback. Furthermore, all experimental results have been replicated three times independently, and all LLM-generated code was verified and tested for correctness by at least two authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility Statement

The appendices provide full proofs, model parameters, dataset specifications, and complete numerical results. The anonymized supplementary source package (attached as the CRAFT-GBM_Supplementary_Source_Code.zip file in the submission) includes model implementations, data acquisition and generation scripts, fixed experimental settings, and benchmark and analysis code. Dependency versions and execution instructions are provided to support reproduction of the experiments. While we expect the relative performance between competitor models to be fully reproduced, the absolute training times may vary depending on the hardware and execution environment.

REFERENCES

Róbert Busa-Fekete and Balázs Kégl. Accelerating adaboost using ucb. In *JMLR Workshop and Conference Proceedings*, volume 7, pp. 111–122, 2009. URL https://proceedings.mlr.press/v7/busa09.html.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785. URL https://arxiv.org/abs/1603.02754v3.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006. URL https://www.jmlr.org/papers/v7/evendar06a.html.

Figshare. Bosch production line performance: Numeric training table. Figshare, version 4, file 37613486, 2022. URL https://doi.org/10.6084/m9.figshare.21118804.v4.

Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1): 1–67, 1991. doi: 10.1214/aos/1176347963. URL https://doi.org/10.1214/aos/1176347963.

Jeremy Goldwasser and Giles Hooker. Statistical significance of feature importance rankings. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1476–1496. PMLR, 2025. URL https://proceedings.mlr.press/v286/goldwasser25a.html.

Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, pp. 507–520. Curran Associates, Inc., 2022. doi: 10.52202/068431-0037. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321. URL https://arxiv.org/abs/1808.03204v8.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991. URL https://arxiv.org/abs/1810.08240v9.

Shibal Ibrahim, Kayhan Behdin, and Rahul Mazumder. End-to-end feature selection approach for learning skinny trees. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2863–2871. PMLR, 2024. URL https://proceedings.mlr.press/v238/ibrahim24a.html.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In John Langford and Joelle Pineau (eds.), *Proceedings of the 29th International Conference on Machine Learning*, pp. 655–662. Omnipress, 2012. URL https://icml.cc/2012/papers/359.pdf.

KDD Cup. KDD Cup 2009: Customer relationship prediction, small dataset, 2009. URL https://www.kdd.org/kdd-cup/view/kdd-cup-2009/Data. Appetency task.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

LIBSVM Dataset Repository. Epsilon (normalized) binary-classification benchmark, 2026. URL https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#epsilon. Exact training and test artifacts accessed 2026-09-03.

Oded Maron and Andrew W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In *Advances in Neural Information Processing Systems*, volume 6, pp. 59–66. Morgan-Kaufmann, 1993. URL https://proceedings.neurips.cc/paper_files/paper/1993/file/02a32ad2669e6fe298e607fe7cc0e1a0-Paper.pdf.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems*, volume 36, pp. 76336–76369. Curran Associates, Inc., 2023. doi: 10.52202/075280-3337. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f06d5ebd4ff40b40dd97e30cee632123-Abstract.html.

Marco Mussi, Alessandro Montenegro, Francesco Trovò, Marcello Restelli, and Alberto Maria Metelli. Best arm identification for stochastic rising bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36953–36989. PMLR, 2024. URL https://proceedings.mlr.press/v235/mussi24b.html.

Bitya Neuhof and Yuval Benjamini. Confident feature ranking. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1468–1476. PMLR, 2024. URL https://proceedings.mlr.press/v238/neuhof24a.html.

OpenML. Jannis: Dataset 41168, version 1, and task 168792, 2026. URL https://www.openml.org/t/168792. Dataset: https://www.openml.org/d/41168; task: https://www.openml.org/t/168792; exact artifacts accessed 2026-09-03.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Mo Tiwari, Ryan Kang, Jaeyong Lee, Chris Piech, Ilan Shomorony, Sebastian Thrun, and Martin J Zhang. MABSplit: Faster forest training using multi-armed bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1223–1237. Curran Associates, Inc., 2022. doi: 10.52202/068431-0090. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/08857467641ad82f635023d530605b4c-Abstract-Conference.html.

UCI Machine Learning Repository. APS Failure at Scania Trucks. Dataset 421, 2023a. URL https://doi.org/10.24432/C51S51.

UCI Machine Learning Repository. Relative location of CT slices on the axial axis. Dataset 206, 2023b. URL https://doi.org/10.24432/C5CP6G.

UCI Machine Learning Repository. Gisette. Dataset 170, 2023c. URL https://doi.org/10.24432/C5HP5B.

UCI Machine Learning Repository. Human activity recognition using smartphones. Dataset 240, 2023d. URL https://doi.org/10.24432/C54S4K.

UCI Machine Learning Repository. YearPredictionMSD. Dataset 203, 2023e. URL https://doi.org/10.24432/C50K61.

Daniel Whiteson. HIGGS. UCI Machine Learning Repository, 2014. URL https://doi.org/10.24432/C5V312.

A Properties of Feature Gain Transformation (FGT)

Proof. The claims in Section 2.1 follow directly from the definition in Equation (2).

Boundedness. For $g \geq 0$ and $s > 0$,

$$0 \leq \frac{g}{g+s} < 1.$$

When $s = \perp$, $\phi(g; \perp) = \mathbf{1}\{g > 0\} \in \{0, 1\}$. Hence every summand in $R_{j,t}$ lies in $[0, 1]$, and averaging over the M output trees gives $R_{j,t} \in [0, 1]$.

Monotonicity and positive-rescaling invariance. For a finite positive scale s,

$$\frac{\partial}{\partial g}\frac{g}{g+s} = \frac{s}{(g+s)^2} > 0,$$

so the transformation $\phi(g; s)$ is strictly increasing in g. For every $a > 0$,

$$\phi(ag; as) = \frac{ag}{ag+as} = \frac{g}{g+s} = \phi(g; s).$$

Thus a common positive rescaling of the gain g and its scale s leaves the transformed value $\phi(g; s)$ unchanged.

Predictability of the scale. By construction, $s_{m,t}$ is committed before accepted round t and is therefore $\mathcal{F}_{t-1}$-measurable. The current gains $\{G_{j,m,t} : j \in S_t\}$ may update only $s_{m,t+1}$, after every output tree in round t succeeds. Consequently, a current gain $G_{j,m,t}$ cannot affect its own same-round FGT scale $s_{m,t}$. □

B Empirical-Bernstein Boundary and Fixed Scale

Fixed scale. After native preprocessing fixes p, fix $n_{\mathrm{ref}} > 0$ (default 300) and choose $z > 0$ and $\rho > 0$ by

$$z - \log(1+z) = 2\log(p/\delta), \qquad \rho = \frac{n_{\mathrm{ref}}}{4z}. \tag{15}$$

The left side increases strictly from zero to infinity, so z exists uniquely. This *normal-reference scale* ρ uses the normal-mixture calibration convention of Howard et al. (2021, Proposition 3) at the reference variance $n_{\mathrm{ref}}/4$. It is a fixed design choice, not an optimum for the gamma-exponential boundary $u_\rho(v; \alpha)$. The value $n_{\mathrm{ref}}/4$ is not an upper bound on the observed squared-error sum $V_{j,t}$. Coverage holds for every fixed positive ρ. The reference n_{ref} may be configured before training, but is never chosen from observed gains, losses, or stopping times. The stopping constants $\epsilon = 0.001$, $P = 10$, and their derived $n_\star = 10000$ are separate from this confidence calibration.

Boundary. For $0 \leq \lambda < 1$, write $\psi_E(\lambda) = -\log(1-\lambda) - \lambda$. Let $P(a, x) = \gamma(a, x)/\Gamma(a)$ be the regularized lower incomplete gamma function. Mix $\exp\{\lambda s - \psi_E(\lambda)v\}$ against the density

$$f_\rho(\lambda) = \frac{\rho^\rho}{\Gamma(\rho)P(\rho, \rho)}(1-\lambda)^{\rho-1}e^{-\rho(1-\lambda)}, \qquad 0 < \lambda < 1.$$

This is the distribution of $1 - G$ conditional on $G < 1$, for a gamma variable G with shape and rate ρ. For every real s and $v \geq 0$, define

$$M_\rho(s, v) = \int_0^1 e^{\lambda s - \psi_E(\lambda)v} f_\rho(\lambda)\, d\lambda.$$

For $s \geq 0$, the domain needed for boundary inversion, integration gives

$$M_\rho(s, v) = \frac{\rho^\rho \Gamma(v+\rho)P(v+\rho, s+v+\rho)e^{s+v}}{\Gamma(\rho)P(\rho, \rho)(s+v+\rho)^{v+\rho}}. \tag{16}$$

For $v \geq 0$ and $\alpha \in (0, 1)$, define $u_\rho(v; \alpha)$ as the unique nonnegative solution of $M_\rho(u_\rho(v; \alpha), v) = 1/\alpha$. Uniqueness follows because the mixture $M_\rho(s, v)$ is continuous and strictly increasing in s, starts at a value of at most one at $s = 0$, and diverges as s increases. The boundary $u_\rho(v; \alpha)$ is nondecreasing in v, since every mixture integrand decreases with v.

Numerical evaluation. The implementation evaluates the log mixture $\log M_\rho(s, v)$ and brackets its root, returning the upper endpoint with an outward numerical margin. It caches boundaries $u_\rho(v^+; \alpha)$ on an upward-rounded logarithmic grid for v: writing $e = \lfloor \log_2(1 + v) \rfloor$, use $\Delta = 2^e/128$ and $v^+ = \lceil (1 + v)/\Delta \rceil \Delta - 1 \geq v$. Monotonicity makes $u_\rho(v^+; \alpha)$ conservative for the exact boundary $u_\rho(v; \alpha)$. Division still uses the current N, including when a new observation adds zero squared error. Unsupported numerical inputs return the full interval $[0, 1]$ rather than a narrowed bound. Numerical tests compare the implementation against high-precision roots. These tests are not a formal certification of floating-point special functions. All coverage and selection guarantees also hold with any conservative numerical radius $\widetilde{r} \geq r$ in place of r, provided the endpoints $L_{j,t}, U_{j,t}$, interval point estimates $c_{j,t}$, half-widths $h_{j,t}$, and selected sets $S_t, S_t^{(f)}$ are all computed from $\widetilde{r}$. The selection-gap bounds then use these implemented radii $\widetilde{r}$ and half-widths $h_{j,t}$.

C Proof of Theorem 1

Proof. Condition throughout on $\mathcal{F}_1$. Under the resulting conditional law, p, $\alpha = \delta/(2p)$, and $\rho > 0$ are fixed.

Predictable squared-error control. Fix feature j, suppress its index, and put $\widehat{Y}_t = I_t a_t$, $D_t = Y_t - \nu_t$, and $m_t = \widehat{Y}_t - \nu_t$. Both $\widehat{Y}_t$ and m_t are predictable, $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = 0$, and $D_t - m_t = Y_t - \widehat{Y}_t \in [-1, 1]$. The empirical-Bernstein exponential inequality used in Howard et al. (2021, Theorem 4 and Appendix A.8) gives, for $0 \leq \lambda < 1$ and $x \geq -1$,

$$e^{\lambda x - \psi_E(\lambda) x^2} \leq 1 + \lambda x.$$

Consequently,

$$\mathbb{E}\Big[e^{\lambda D_t - \psi_E(\lambda)(Y_t - \widehat{Y}_t)^2} \mid \mathcal{F}_{t-1}\Big] \leq e^{\lambda m_t}(1 - \lambda m_t) \leq 1.$$

Since $V_u = \sum_{t=2}^{u} (Y_t - \widehat{Y}_t)^2$, multiplication over rounds makes $\exp\{\lambda S_u - \psi_E(\lambda) V_u\}$ a nonnegative supermartingale starting at one, where $S_u = \sum_{t=2}^{u} D_t$. Applying the same argument to $-Y_t$ and $-\widehat{Y}_t$ gives the corresponding supermartingale for $-S_u$, with exactly the same V_u.

Mixture, inversion, and joint coverage. Integration against the fixed density f_ρ preserves the supermartingale property by conditional Tonelli's theorem. Ville's inequality therefore bounds the probability that either $M_\rho(S_u, V_u)$ or $M_\rho(-S_u, V_u)$ ever reaches $1/\alpha$ by at most 2α. Monotonicity in the first argument yields $|S_u| < u_\rho(V_u; \alpha)$ simultaneously at all rounds outside these events. For $N_u > 0$, the identity

$$S_u = A_u - \sum_{t=2}^{u} \nu_t = N_u(\widehat{\mu}_u - \overline{\nu}_u)$$

gives the radius r_u in Equation (6). A conditional union bound over the fixed p features has total error at most $2p\alpha = \delta$. Random observation counts N_u and adaptive stopping cause no additional error allocation because the event already covers every round. No independence across features, rounds, native EFB groups, GOSS draws, or multiclass outputs is used. □

D Proofs for Feature Selection

Relation to feature-rank inference. Statistical racing uses uncertainty to allocate observations adaptively (Maron & Moore, 1993; Busa-Fekete & Kégl, 2009). MABSplit allocates samples to candidate splits within a tree node (Tiwari et al., 2022), whereas CRAFT-GBM accumulates root-gain evidence across boosting rounds. Simultaneous rank intervals yield confidence sets containing top-ranked features (Neuhof & Benjamini, 2024), while sequential sampling procedures can verify feature-importance rankings (Goldwasser & Hooker, 2025). These guarantees concern feature-importance targets under their respective sampling assumptions. Here, joint coverage of accumulated expected gains $\overline{\nu}_{j,t-1}$ implies possible-best retention for the Endogenous rule and an approximate top-K gap bound for the Exogenous rule. The latter does not assert exact recovery of the top-K set.

Proof of Proposition 1. Write $C_t = \max_c L_{c,t}$.

Suppose first that $U_{j,t} < C_t$. Choose c with $L_{c,t} = C_t$. Every $v \in \mathcal{C}_t$ satisfies

$$v_j \le U_{j,t} < L_{c,t} \le v_c,$$

so feature j cannot maximize any parameter vector v in the rectangular confidence region $\mathcal{C}_t$.

Conversely, suppose $U_{j,t} \ge C_t$. Define $v \in \mathbb{R}^p$ by

$$v_j = U_{j,t}, \qquad v_c = L_{c,t} \quad (c \ne j).$$

The intervals $[L_{j,t}, U_{j,t}]$ are nonempty, so $v_j \in [L_{j,t}, U_{j,t}]$, and all other coordinates also lie in their intervals $[L_{c,t}, U_{c,t}]$. Moreover, $v_c = L_{c,t} \le C_t \le U_{j,t} = v_j$ for every $c \ne j$. Thus $v \in \mathcal{C}_t$ and feature j is a maximizer of v. The two sets are therefore equal. Since every member of S_t is a maximizer for at least one allowed vector v, deleting any member would remove a possible-best feature. This proves the stated sharpness relative to $\mathcal{C}_t$. □

Proof of Theorem 2. Let Ω_δ be the joint time-uniform coverage event Ω_δ from Theorem 1. This event has conditional probability at least $1 - \delta$. Fix any boosting round $t \ge 21$ and work on Ω_δ. Round 2 selected every feature, so $N_{j,t-1} > 0$ for all j.

A useful bounded-interval sandwich. Coverage gives $|\widehat{\mu}_{j,t-1} - \overline{\nu}_{j,t-1}| \le r_{j,t-1}$, and $\overline{\nu}_{j,t-1} \in [0, 1]$. From the definition of the bounded endpoints $L_{j,t}, U_{j,t}$,

$$L_{j,t} \le \overline{\nu}_{j,t-1} \le U_{j,t}, \qquad L_{j,t} \ge \overline{\nu}_{j,t-1} - 2r_{j,t-1}, \qquad U_{j,t} \le \overline{\nu}_{j,t-1} + 2r_{j,t-1}. \tag{17}$$

Indeed, $\widehat{\mu}_{j,t-1} - r_{j,t-1} \le \overline{\nu}_{j,t-1}$ and $0 \le \overline{\nu}_{j,t-1}$, which proves the first lower-endpoint inequality. The other coverage inequality and $\overline{\nu}_{j,t-1} \le 1$ give the upper-endpoint inequality. Also, $\widehat{\mu}_{j,t-1} \ge \overline{\nu}_{j,t-1} - r_{j,t-1}$ implies $L_{j,t} \ge \overline{\nu}_{j,t-1} - 2r_{j,t-1}$, while $\widehat{\mu}_{j,t-1} \le \overline{\nu}_{j,t-1} + r_{j,t-1}$ implies $U_{j,t} \le \overline{\nu}_{j,t-1} + 2r_{j,t-1}$.

1. No true maximizer is omitted. Let $j^\star \in \mathcal{M}_t^\star$. By Equation (17), $U_{j^\star,t} \ge \overline{\nu}_{j^\star,t-1} = \nu_t^\star$. For every feature c, $L_{c,t} \le \overline{\nu}_{c,t-1} \le \nu_t^\star$, so $C_t = \max_c L_{c,t} \le \nu_t^\star$. Hence $U_{j^\star,t} \ge C_t$, and $j^\star \in S_t$. Since this holds for every true maximizer $j^\star \in \mathcal{M}_t^\star$,

$$\mathcal{M}_t^\star \subseteq S_t, \qquad K_t \ge |\mathcal{M}_t^\star|.$$

2. Every omitted feature is strictly suboptimal. If $j \notin S_t$, then $U_{j,t} < C_t$. Choose c attaining $C_t = L_{c,t}$. Since $U_{c,t} \ge L_{c,t} = C_t$, feature c is retained. Equation (17) then gives

$$\overline{\nu}_{j,t-1} \le U_{j,t} < C_t = L_{c,t} \le \overline{\nu}_{c,t-1},$$

which is the claimed strict comparison.

3. Approximate-best guarantee for each retained feature. Choose $j^\star \in \mathcal{M}_t^\star$ such that $r_{j^\star,t-1} = r_t^\star$. If $j \in S_t$, then

$$U_{j,t} \ge C_t \ge L_{j^\star,t} \ge \nu_t^\star - 2r_t^\star,$$

where the final inequality is Equation (17). The same equation also gives $U_{j,t} \le \overline{\nu}_{j,t-1} + 2r_{j,t-1}$. Combining the two displays yields

$$0 \le \nu_t^\star - \overline{\nu}_{j,t-1} \le 2r_{j,t-1} + 2r_t^\star,$$

which is Equation (9). Summing over $j \in S_t$ and dividing by K_t gives

$$0 \le \nu_t^\star - \frac{1}{K_t} \sum_{j \in S_t} \overline{\nu}_{j,t-1} \le 2r_t^\star + \frac{2}{K_t} \sum_{j \in S_t} r_{j,t-1}.$$

The minimum-radius maximizer $j^\star$ is retained by part 1, so $r_t^\star \le \max_{j \in S_t} r_{j,t-1}$, and the selected-set average of the radii $r_{j,t-1}$ is bounded by the same maximum. Therefore the right-hand side is at most $4 \max_{j \in S_t} r_{j,t-1}$, proving Equation (10).

4. Exact recovery. If $K_t = 1$, part 1 places the nonempty maximizer set $\mathcal{M}_t^\star$ inside a singleton, so the sole retained feature is the unique maximizer. If $\mathcal{M}_t^\star = \{1, \dots, p\}$, part 1 directly gives $S_t = \mathcal{M}_t^\star$.

Now suppose $\mathcal{M}_t^\star \neq \{1, \dots, p\}$ and the separation condition in Equation (11) holds. Let

$$L_t^\star = \max_{c \in \mathcal{M}_t^\star} L_{c,t}.$$

For every nonmaximizer $j \notin \mathcal{M}_t^\star$, $L_{j,t} \leq U_{j,t} < L_t^\star$. Hence the global lower-endpoint maximum is $C_t = L_t^\star$, and every nonmaximizer satisfies $U_{j,t} < C_t$ and is omitted. Every maximizer $j \in \mathcal{M}_t^\star$ is retained by part 1, so $S_t = \mathcal{M}_t^\star$.

All four conclusions hold for every boosting round $t \geq 21$ on the single event Ω_δ, which completes the proof. □

Proof of Corollary 1. Fix a round covered by Theorem 2 and work on the same joint coverage event Ω_δ. By optimality of $S_t^{\star,K_t}$,

$$\sum_{j \in S_t^{\star,K_t}} \overline{\nu}_{j,t-1} \geq \sum_{j \in S_t} \overline{\nu}_{j,t-1},$$

which proves the lower bound in Equation (12).

The selection rule retains every feature with $U_{j,t} \geq C_t$ and omits only features with $U_{j,t} < C_t$. Therefore every selected upper endpoint $U_{j,t}$, $j \in S_t$, is strictly larger than every omitted upper endpoint $U_{j,t}$, $j \notin S_t$. All ties at the boundary C_t are retained. Let

$$A = S_t^{\star,K_t} \setminus S_t, \qquad B = S_t \setminus S_t^{\star,K_t}.$$

The two sets A, B have the same size. If they are empty, the result is immediate. Otherwise choose any bijection $\pi : A \to B$. For each $j \in A$, feature j is omitted and $\pi(j)$ is selected, so coverage and Equation (17) imply

$$\overline{\nu}_{j,t-1} \leq U_{j,t} < U_{\pi(j),t} \leq \overline{\nu}_{\pi(j),t-1} + 2r_{\pi(j),t-1}.$$

Since the common elements of $S_t^{\star,K_t}$ and S_t cancel,

$$\begin{aligned}
&\sum_{j \in S_t^{\star,K_t}} \overline{\nu}_{j,t-1} - \sum_{j \in S_t} \overline{\nu}_{j,t-1} \\
&\quad = \sum_{j \in A} \left(\overline{\nu}_{j,t-1} - \overline{\nu}_{\pi(j),t-1} \right) \\
&\quad \leq 2 \sum_{i \in B} r_{i,t-1} \leq 2K_t \max_{i \in S_t} r_{i,t-1}.
\end{aligned}$$

Dividing by K_t proves the upper bound in Equation (12). The set S_t and its selection size K_t are measurable with respect to the committed past $\mathcal{F}_{t-1}$, and this exchange argument is deterministic on the already joint event Ω_δ. No further union bound over realized values of K_t is required. □

Proof of Theorem 3. Work on the joint event Ω_δ from Theorem 1. Fix $t \geq 3$, and write

$$A = S_t^{\star,K_f} \setminus S_t^{(f)}, \qquad B = S_t^{(f)} \setminus S_t^{\star,K_f}.$$

These sets A, B have equal size. Choose any bijection $\pi : A \to B$. Because $S_t^{(f)}$ contains the K_f largest interval point estimates $c_{j,t}$, $c_{j,t} \leq c_{\pi(j),t}$ for every $j \in A$. Coverage and the known target range $[0, 1]$ imply $\overline{\nu}_{j,t-1} \in [c_{j,t} - h_{j,t}, c_{j,t} + h_{j,t}]$, so

$$\overline{\nu}_{j,t-1} - \overline{\nu}_{\pi(j),t-1} \leq c_{j,t} + h_{j,t} - c_{\pi(j),t} + h_{\pi(j),t} \leq h_{j,t} + h_{\pi(j),t}.$$

Canceling common members and summing proves the middle bound in Equation (14). Since $|A| = |B| \leq K_f$ and $0 \leq h_{j,t} \leq 1/2$, the remaining upper bounds follow. The lower bound follows from the optimality of $S_t^{\star,K_f}$. The argument is deterministic on the joint coverage event Ω_δ and applies at every selection round, so no further union bound over rounds or selected sets $S_t^{(f)}$ is needed. □

E TRAINING-LOSS STOPPING AND RETURNED ENDPOINT

The RQ1 endogenous controller uses only the complete training set. There is no stopping-validation set, and final-test data are unavailable to it. Let F_0 be the objective's training-derived initial predictor F_0 before any boosting tree, and let F_t be the predictor F_t after t successfully completed boosting rounds. Write L_t for its prespecified empirical training loss L_t: native binary or multiclass log loss, or RMSE for regression. The objective transform, weights, and aggregation remain fixed.

Evaluate L_0 once. The protocol rejects $L_0 = 0$ as a degenerate case rather than returning a zero-tree model. For finite $L_0 > 0$, set

$$\ell_t = \frac{L_t}{L_0}, \qquad \epsilon = 0.001, \qquad P = 10.$$

Thus $\ell_0 = 1$, and the fixed raw-scale improvement threshold η is $\eta = \epsilon L_0$, not a fraction of a moving incumbent. Nonfinite or negative losses L_t, or a nonpositive or nonfinite computed threshold η, are failures rather than plateau evidence.

The stopping controller starts after round 2, even during the endogenous full-feature initialization. Set

$$a_2 = \min\{\ell_0, \ell_1, \ell_2\}, \qquad h_2 = 2.$$

The anchor a_t controls clock resets, not checkpoint selection. For every completed $t \geq 3$, first update

$$(a_t, h_t) = \begin{cases} (\ell_t, t), & \ell_t < a_{t-1} - \epsilon, \\ (a_{t-1}, h_{t-1}), & \text{otherwise}, \end{cases} \tag{18}$$

and only then test $t - h_t \geq P$. Equality at the threshold $a_{t-1} - \epsilon$ does not reset the clock. The stopping round τ is

$$\tau = \inf\{t \geq 3 : t - h_t \geq P\}. \tag{19}$$

The measured endogenous fit returns F_τ, not the loss minimizer over earlier rounds. In RQ1, its untimed pre-run reports the round count $t^{(0)}$ and average selected feature fraction $\bar{f}$ but is discarded without test prediction. No external round cap is supplied to an endogenous fit. Fixed-horizon comparators instead return the model at their prescribed round count. Passive patience bookkeeping does not shorten those fits.

Theorem 4 (Finite training-loss stopping)**.** *Assume L_0 is finite and positive and native training can continue while producing a finite nonnegative training-loss sequence (L_t). For $\epsilon \in (0, 1)$, integer $P \geq 1$, and Equations* (18)–(19)*:*

1. *τ is a stopping time for the completed-round training history $\mathcal{F}_t$.*
2. $$\tau \leq 2 + P \left\lceil \frac{1}{\epsilon} \right\rceil;$$
3. *for every completed $t \geq 2$,* $$0 \leq a_t - \min_{0 \leq s \leq t} \ell_s \leq \epsilon;$$
4. *at a patience stop, $\ell_s \geq a_\tau - \epsilon$ for $s = \tau - P + 1, \ldots, \tau$.*

For the fixed $\epsilon = 0.001$ and $P = 10$, the deterministic upper bound is 10,002 completed rounds.

Proof. Every update and decision uses only the completed-round history $\mathcal{F}_t$, proving the stopping-time statement. Initially $0 \leq a_2 \leq 1$, and every reset decreases the anchor a_t by strictly more than ϵ, so if J resets occur then $J < 1/\epsilon$, hence $J + 1 \leq \lceil 1/\epsilon \rceil$. There are at most P completed rounds before the first reset, between successive resets, and between the final reset and stopping because a qualifying improvement on the P-th round is processed first. This gives the finite bound. Induction on Equation (18) gives the anchor inequality. During the final P rounds no reset occurs and the anchor a_t is constant, which gives the last assertion. □

The bound is neither a recommended training length nor a certificate that later unevaluated rounds cannot improve. The anchor inequality concerns the internal anchor a_t, not the returned model's loss L_τ. The $n_\star = P\lceil 1/\epsilon \rceil = 10000$ term describes the stopping rule, not the separate confidence reference $n_{\mathrm{ref}} = 300$. Training is not instructed to reach either value. The implementation checks the 10,002-round limit as an integrity condition after testing patience, not as an independently chosen training budget.

Because Theorem 1 is joint over all completed rounds, its feature statements remain valid through τ without another error allocation. If native training cannot complete another round before the required endpoint, the fit is unavailable rather than a successful patience stop. A completed round that reaches patience or a prescribed fixed endpoint is accepted. Cancellation, timeout, and numerical failure remain separate from normal stopping. Predictive quality is evaluated on the held-out test set after the endpoint models are frozen.

F Algorithm and Execution Details

At the beginning of training, after native Dataset construction has fixed the feature map and revealed the feature count p, CRAFT-GBM computes the per-tail error allocation α and the positive mixture scale ρ defined in Appendix B. Both are frozen before round 1. The calibration uses the prespecified reference count n_{ref}, not the stopping tolerance ϵ, patience P, requested or realized training length, observed gains, or losses. The fixed feature map, native EFB bundling map, and storage representation are included in $\mathcal{F}_1$ and are not recalibrated during training. RQ1 endogenous fits receive no external round limit, while fixed-horizon fits receive the endpoint prescribed by their experimental protocol. Neither endpoint enters the mixture scale ρ.

Round 1 initializes scales s_m only, and round 2 supplies the first confidence observation $R_{j,2}$ for every feature. Endogenous fits remain full-feature through round 20 if reached, while updating statistics A_j, N_j, V_j normally. A selected feature with zero gain still increments its observation count N_j, whereas an omitted feature changes none of A, N, V. The median $\mathrm{median}(H_m)$ uses the midpoint for an even number of values. The training-loss controller in Appendix E accepts the patience-triggering round τ and returns that model F_τ. Fixed-horizon fits return their prescribed endpoint. An incomplete attempt is not a successful stop.

Algorithm 1 CRAFT-GBM

Inputs: fixed features $[p] = \{1, \dots, p\}$ and canonical identifiers, error level $\delta \in (0, 1)$, fraction $f \in (0, 1]$, native learner, and endpoint controller.
Output: model at the controller's accepted endpoint.

1. **Initialize:** $A_j = N_j = V_j = 0$ for $j \in [p]$, $s_m = \perp$ for $m \in [M]$, and $\alpha = \delta/(2p)$. Freeze ρ by Equation (15). Set $K_f = \max\{1, \lfloor fp + 1/2 \rfloor\}$ and $t = 1$.
2. **While** the controller requests round t:
   (a) **Select from committed history.** For $t \geq 3$, compute (L_j, U_j) by Equations (6) and (7), and set the interval point estimate $c_j = (L_j + U_j)/2$. Set

$$S = \begin{cases} [p], & t \leq 2 \text{ or } (f = 1,\ t \leq 20), \\ \{j : U_j \geq \max_c L_c\}, & t \geq 21,\ f = 1, \\ \mathrm{First}_{K_f}\big(\mathrm{sort}_{j \in [p]}(-c_j, N_j, \mathrm{id}(j))\big), & t \geq 3,\ f < 1. \end{cases}$$

   Here $\mathrm{id}(j)$ is the canonical feature identifier $\mathrm{id}(j)$. Sorting is ascending lexicographic, and First_{K_f} returns the first K_f feature indices.
   (b) **Build a tentative round.** Fix S before GOSS and use it for all M output trees. Record $G_{j,m,t}$ from Equation (1) for every $j \in S$ and $m \in [M]$.
   (c) **If the native round is discarded or an exception occurs:** abort pending selector updates and restore its committed mask, leaving statistics A_j, N_j, V_j, scales s_m, and the endpoint controller unchanged. Do not advance t. Native code handles model termination, and exceptions fail the fit rather than resuming from a partially built round.
   (d) **Otherwise, stage the updates.** For $t \geq 2$ and $j \in S$, using the pre-round scales s_m and pre-update statistics A_j, N_j, V_j, compute

$$R_j = \frac{1}{M} \sum_{m=1}^{M} \phi(G_{j,m,t}; s_m), \qquad a_j = \frac{1/2 + A_j}{1 + N_j},$$

$$(A_j^+, N_j^+, V_j^+) = (A_j + R_j,\ N_j + 1,\ V_j + (R_j - a_j)^2).$$

   For $t = 1$ or $j \notin S$, set $(A_j^+, N_j^+, V_j^+) = (A_j, N_j, V_j)$. For each output m, let H_m be the multiset of positive $G_{j,m,t}$ values, one per selected feature $j \in S$, and set

$$s_m^+ = \begin{cases} \mathrm{median}(H_m), & H_m \neq \varnothing, \\ s_m, & H_m = \varnothing. \end{cases}$$

   (e) **Commit** the completed model F_t, $S_t = S$, $K_t = |S|$, and $(A, N, V, s) \leftarrow (A^+, N^+, V^+, s^+)$. Update the endpoint controller from this round and set $t \leftarrow t + 1$.

G NATIVE EXECUTION

Native LightGBM constructs EFB groups before boosting (Ke et al., 2017). Loading a binary Dataset restores the same fixed representation. Let $\mathcal{E} = \{1, \ldots, B\}$ index these groups and let $D_b \subseteq \{1, \ldots, p\}$ contain the valid features stored in group b. For any selected feature set S, define its direct EFB-group closure $\mathcal{E}(S)$

$$\mathcal{E}(S) = \{b \in \mathcal{E} : D_b \cap S \neq \varnothing\}. \tag{20}$$

In the validated forced-column-wise CPU path, a group is invoked exactly when it belongs to this closure $\mathcal{E}(S)$. The closure $\mathcal{E}(S)$ is nevertheless a coarse work descriptor: an ordinary packed group constructs its shared histogram when any member is enabled, and the unchanged multi-value subcolumn heuristic may process all member columns in an implicated group. This physical processing does not make features outside the exact mask S eligible or update their statistics. CRAFT-GBM neither rebuilds EFB nor changes native binning, packed storage, subset construction, or histogram arithmetic.

Native GOSS selects and reweights rows (Ke et al., 2017). The selected feature mask S and its direct group closure $\mathcal{E}(S)$ are fixed before the current row draw. Rows and weights then determine root gains $G_{j,m,t}$. No per-feature row sample or additional histogram pass is introduced.

H MODEL PARAMETERS OF CRAFT-GBM

Table 1: Model parameters and conventions of CRAFT-GBM

Quantity	Value	Role and limitation
Enable selector	`true` for CRAFT-GBM	Set `adaptive_feature_selection=true`. The native default is `false`.
Feature count	p	Number of features retained by native preprocessing, fixed before covered gains.
Joint-confidence error level	$\delta = 0.05$ (default)	Joint-confidence error level δ over the fixed features for all boosting rounds, fixed at the beginning of training.
Per-tail error level	$\alpha = \delta/(2p)$	Derived from the fixed feature count p and joint-confidence error level δ, not from observed gains.
Boosting-round endpoint	endogenous or protocol-fixed	RQ1 endogenous fits use training-loss patience P without an external cap. Fixed-horizon fits use their prescribed endpoint. Neither determines the feature-confidence boundary $u_\rho(v; \alpha)$.
`feature_fraction`	1.0 (default) or $f \in (0, 1)$	With the selector enabled, 1.0 invokes the endogenous rule for the selection size K_t, while $f < 1$ invokes Exogenous K_f as defined in Equation (13).
Confidence reference	$n_{\mathrm{ref}} = 300$ (default)	A scale reference n_{ref} configurable before training, not a training horizon.
Mixture scale	$\rho = n_{\mathrm{ref}}/(4z)$	Derived by Equation (15) and frozen before round 1. The training horizon, gains, and losses are excluded.
Stopping rule	$\epsilon = 0.001,\ P = 10,\ n_\star = 10000$	Fixed training-loss improvement tolerance ϵ and patience P. The derived $n_\star = P\lceil 1/\epsilon \rceil = 10000$ is their deterministic stopping-bound term, not a coverage horizon or execution target.

Table 1 lists the model parameters and conventions. The experiments fix $n_{\mathrm{ref}} = 300$ in RQ1 and 1000 in RQ2.

Supported execution. The extension supports fresh, continuous, single-machine fits with native objectives and `boosting=gbdt`, `device_type=cpu`, `tree_learner=serial`, `force_col_wise=true`, `force_row_wise=false`, `enable_bundle=true`, `is_enable_sparse=true`, and `feature_fraction_bynode=1.0`. Row sampling is native GOSS or no sampling, with ordinary and query bagging disabled. Unsupported configurations include custom objectives, GPU or distributed training, warm starts, native early stopping, linear trees, quantized gradients, extra trees, CEGB, monotone or interaction constraints, feature contributions, and forced splits.

The native extension implements feature selection and confidence updates. The supplied benchmark driver separately implements the RQ1 training-loss stopping rule: it updates the booster one round at a time, evaluates training loss, applies patience, and saves the completed endpoint. Enabling the selector alone does not install this stopping rule.

The joint-confidence error level δ may be any $\mathcal{F}_1$-measurable value fixed at the beginning of training, but changing it after training begins to obtain a preferred selection size K_t invalidates the stated guarantee. Each interval width $2h_{j,t}$ uses the fixed feature count p and that feature's actual selection history. The feature-selector controls are δ (`adaptive_feature_confidence_delta`), native `feature_fraction` f, and the confidence reference n_{ref} `adaptive_feature_calibration_reference_count`. The feature count p follows from native construction. The fraction f is fixed for the fit. Training-time reconfiguration is not a supported way to choose a preferred selection size K_t. The internal ϵ, P values and the fixed mixture-scale calibration convention are disclosed design choices rather than parameter-free scientific consequences.

I SYNTHETIC DATA GENERATION

The four synthetic families complement the public datasets with controlled signal and input structures. Every generated replicate has 2,000 numerical features and 100,000 training, 20,000 validation, and 100,000 test observations. The release contains ten replicates per family, indexed 00–09, while the publication benchmark retains the prespecified IDs 00–04, chosen by index before model results were inspected. All models share the same observed arrays within a replicate and use model seed 314159. Replicates are separate learning problems and are never concatenated or replaced according to their outcomes. The generated validation splits are unused, and RQ2 instead creates its inner validation split from the training file. All four families enter RQ1 and RQ2.

Below, coordinates refer to the canonical representation before a replicate-specific random column permutation, shared across its three splits. All stored inputs are float32: Sparse 1/2 use sparse CSR matrices and Dense 1/2 use dense arrays. Binary labels are uint8, whereas Dense 1 labels and all supplied oracle targets are float64. Native preprocessing is fitted on training data only. Sparse zeros remain observed values, with `zero_as_missing=false`.

Sparse 1: independent sparse Gaussian inputs. For each row and coordinate, independently draw

$$B_j \sim \text{Bernoulli}(a), \qquad Z_j \sim N(0,1), \qquad X_j = B_j Z_j/\sqrt{a}, \qquad a = 0.05.$$

The first 100 canonical coordinates have independent signs $b_j \in \{-1,+1\}$, drawn uniformly once per replicate and held fixed across splits. With $\sigma(z) = (1+e^{-z})^{-1}$, define

$$f_*(X) = \frac{2}{\sqrt{100}} \sum_{j=1}^{100} b_j X_j, \qquad p_*(X) = \sigma(f_*(X)), \qquad Y \mid X \sim \text{Bernoulli}(p_*(X)).$$

The other 1,900 coordinates are nuisance features. The Bernoulli mask B_j is sampled entrywise, giving density a of 0.05 in expectation rather than a fixed nonzero count. Under the ideal real-valued law, each coordinate X_j has variance one, $\text{Var}(f_*(X)) = 4$, and $\Pr(Y=1) = 1/2$. The probability of no active signal coordinate is $0.95^{100} \approx 0.00592$.

Sparse 2: exclusive categorical indicators. Draw 100 independent fields C_g, each uniform on $\{1,\dots,20\}$, and release the numerical indicators $X_{g,c} = \mathbf{1}\{C_g = c\}$. Every row has exactly 100 ones, hence density 0.05. Select five signal fields $\mathcal{G}_*$ uniformly without replacement. For each, randomly permute ten $+1$ and ten -1 category effects $\theta_{g,c}$, once per replicate, and set

$$f_*(X) = \frac{2}{\sqrt{5}} \sum_{g \in \mathcal{G}_*} \theta_{g,C_g}, \qquad p_*(X) = \sigma(f_*(X)), \qquad Y \mid X \sim \text{Bernoulli}(p_*(X)).$$

Thus 100 indicator columns are signal-bearing and 1,900 are nuisance. The five selected effects θ_{g,C_g}, $g \in \mathcal{G}_*$, are independent Rademacher variables, giving score $f_*(X)$ variance four and marginal class probability $\Pr(Y=1)$ one-half. Sparse 1/2 sample rows and splits independently conditional on their fixed replicate metadata, with separate label streams. Their scores $f_*(X)$ and probabilities $p_*(X)$ are evaluated in float64 from the stored float32 inputs X before sampling labels Y. All models receive the same numerical one-hot representation for Sparse 2.

Dense 1: dense nonlinear regression. We implement the Friedman-1 response (Friedman, 1991) with 1,995 additional nuisance coordinates. Draw independent $U_j \sim \text{Uniform}(0,1)$, store X_j as their float32 values, and compute

$$f_*(X) = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5, \qquad Y = f_*(X) + \xi,$$

where $\xi \sim N(0,1)$ is independent of the inputs. Each split uses separate covariate and noise streams. The response function $f_*(X)$ is evaluated in float64 after input rounding, so the saved f_* corresponds to the released inputs. The signal-coordinate map preserves formula order when undoing the column permutation.

Dense 2: dense classification with redundant features. One call to `make_classification` in scikit-learn 1.8.0 (Pedregosa et al., 2011) generates the complete 220,000-row pool for each

replicate. The parameters are 2,000 features, five original informative coordinates, 15 redundant linear combinations, no repeated coordinates, two equally weighted classes, and 16 clusters per class, with `class_sep=1`, `hypercube=True`, `shift=0`, `scale=1`, and `shuffle=False`. The remaining 1,980 coordinates are independent standard-normal nuisance features. The 32 clusters occupy all vertices of $\{-1,+1\}^5$, initially with 6,875 rows per cluster. Within each cluster, standard-normal coordinates undergo a random linear transformation before translation to the vertex, while the redundant coordinates use a shared random linear map. Both types of map have entries drawn independently from $\mathrm{Uniform}(-1,1)$. The pinned implementation assigns class $k \bmod 2$ to vertex k in binary order, $k = 0,\dots,31$, giving the fifth canonical coordinate a class-dependent mean before label randomization. With `flip_y=0.01`, selected labels are reassigned uniformly to either class, so the expected changed-label fraction is 0.005.

A separate outcome-blind row permutation partitions the pool into the fixed training, validation, and test sizes, without stratification or rebalancing. These splits are disjoint partitions of a pool with fixed initial cluster counts, rather than independent unconditional mixture samples. Float32 storage makes the linear redundancy approximate, but all 20 informative or redundant coordinates count as signal-bearing. Dense 2 is MADELON-style, not the original UCI MADELON dataset, and supplies no Bayes-posterior oracle.

Reproducibility. The root seed is 20260909. NumPy streams use PCG64 initialized by `SeedSequence` with key $(\text{family}, \text{replicate}, \text{split}, \text{role})$. Family IDs are 1–4 for Sparse 1, Sparse 2, Dense 1, and Dense 2, and split IDs are 0 for metadata and 1–3 for training, validation, and test. Metadata role 0 determines the column permutation and, for Sparse 1/2, signs b_j or field contrasts $\theta_{g,c}$. Within splits, Sparse 1 uses roles 0/1/2 for masks/values/labels (B_j, Z_j, Y), Sparse 2 uses 0/2 for categories/labels (C_g, Y), and Dense 1 uses 0/1 for covariates/noise (U_j, ξ). Dense 2 derives its scikit-learn integer seed from the first uint32 state value of key $(4, r, 0, 1)$ for replicate r, and uses key $(4, r, 0, 2)$ for the pool-row permutation. Generation pins Python 3.12.14, NumPy 2.3.5, SciPy 1.17.0, and scikit-learn 1.8.0 for Dense 1/2. Sparse 1 and Dense 1 use 1,024-row generation blocks. Release records bind source and environment information to artifact hashes and validation reports, and no replicate is rejected for distributional fluctuations.

Oracle evaluation and scope. Oracle targets and feature-role maps are unavailable to fitting, stopping, and feature-fraction selection. After all scheduled models for the experimental arm are frozen, evaluation uses each replicate's n test inputs. For Sparse 1/2, let q_i be the fitted model's predicted probability q_i at X_i, clipped to $[10^{-15}, 1-10^{-15}]$. We additionally report conditional cross-entropy

$$\frac{1}{n}\sum_{i=1}^{n}\left[-p_*(X_i)\log q_i - (1-p_*(X_i))\log(1-q_i)\right]$$

and mean $\mathrm{KL}(\mathrm{Bernoulli}(p_*(X_i)) \| \mathrm{Bernoulli}(q_i))$. Dense 1 reports direct known-function MSE, $n^{-1}\sum_i(\widehat{f}(X_i) - f_*(X_i))^2$, alongside test RMSE. We do not substitute test MSE minus one for this quantity. These diagnostics concern prediction, not confidence coverage for accumulated transformed root gains $\overline{\nu}_{j,t}$. Sparse 1/2 have additive logits $f_*(X)$, and equal density and score variance do not imply equal Bayes difficulty. Sparse 2's within-field exclusivity does not establish native EFB grouping or a unique minimal indicator support. Cross-family comparisons do not isolate a causal effect of density.

J Experimental Details and Numerical Results

The following tables give all reported dataset–configuration combinations. Unlike the 16-experiment main RQ1 display, these complete results retain both +4 expansions and summarize all 18 measured experiments in RQ1. The main-paper omission is a post-run presentation choice, not an experimental exclusion. Both expansion sizes remain in the appendix. The ten public datasets are HIGGS, Epsilon, Bosch, YearMSD, APS Failure, KDD 2009, Gisette, Jannis, HAR, and CT Scan (Whiteson, 2014; LIBSVM Dataset Repository, 2026; Figshare, 2022; UCI Machine Learning Repository, 2023e;a; KDD Cup, 2009; UCI Machine Learning Repository, 2023c; OpenML, 2026; UCI Machine Learning Repository, 2023d;b). Together with Sparse 1/2 and Dense 1/2 (Appendix I), they form 14 source families. RQ1 adds two expansions each of HAR and YearMSD, giving 18 experiments. Each has five paired measurements: model seeds for nine fixed public splits, patient folds for CT, and generated replicates 00–04 for synthetic and expanded experiments. RQ1 reports 450 completed training/test evaluations. The original RQ2 campaign comprises 1,470 evaluations. RQ2 replaces the original exogenous results with 350 point-estimate-ranking fits and 70 fresh paired Full fits. A further 770 fits supply the two ablations and their Full references. Replaced results are not pooled with their replacements. YearMSD, CT, Dense 1, and the YearMSD expansions are regression experiments, while the others are classification experiments.

Table 2: Public datasets and fixed outer splits. C denotes classification with log loss, and R denotes regression with RMSE. Training counts include any inner validation split used for RQ2 tuning.

Dataset	Task	Training rows	Test rows	Outer split
HIGGS	C	10,500,000	500,000	First/last rows
Epsilon	C	400,000	100,000	Official
Bosch	C	946,997	236,750	Stratified
YearMSD	R	463,715	51,630	First/last rows
APS Failure	C	60,000	16,000	Official
KDD 2009	C	40,000	10,000	Stratified
Gisette	C	6,000	1,000	Official labeled files
Jannis	C	56,102	27,631	OpenML task 168792
HAR	C	7,352	2,947	Subject-disjoint
CT Scan	R	Fold-dependent	Fold-dependent	Patient-disjoint

Table 2 identifies the ten public families. Bosch uses numeric inputs only, KDD 2009 is the Appetency task, and YearMSD denotes YearPredictionMSD. Bosch and KDD use frozen, class-stratified splits ranked by SHA-256 with seed 271828. Gisette uses its official training file for fitting and its labeled file named validation exclusively for final testing. HAR is the smartphone activity dataset, with 21 training and nine test subjects. CT Scan denotes Relative Location of CT Slices and excludes patient IDs from the features. Its five training/test row-count pairs are 42,792/10,708, 42,784/10,716, 42,781/10,719, 42,851/10,649, and 42,792/10,708.

Reproducibility seeds. The nine fixed-split public datasets use measured model seeds 73001–73005. CT fold $i \in \{1, \ldots, 5\}$ uses seed $73000 + i$, with one measurement per fold. Synthetic and expanded experiments use generated replicates 00–04 and measured model seed 314159. Every RQ1 pre-run uses model seed 271828.

The expanded experiments append four or eight independently row-permuted copies of each source feature, giving five or nine times the original width before native preprocessing. Their official train/test row assignments are preserved. The permutations are label-blind but use pooled covariates, and finite-sample split gains of the added columns need not be zero.

Absolute training times and test losses are arithmetic means with sample standard deviations across five measurements. These are not standard errors. CT instead uses the mean and fold range because its training folds overlap. Loss is in its native units: natural-log loss per observation for classification and RMSE for regression. Raw losses are not pooled across datasets. Rounds count boosting iterations, not individual class-specific trees. A retained fraction averages selected feature counts K_s over all rounds, including initialization.

Separate oracle tables report Sparse 1/2 conditional cross-entropy and KL divergence, and Dense 1 error against its known regression function f_*. These use evaluation-only generator information. They do not measure coverage of the feature-gain confidence bounds $[L_{j,t}, U_{j,t}]$.

Paired percentage changes use the geometric mean of the five candidate/reference ratios, followed by $100(r-1)$. They are not ratios of the arithmetic means in the absolute-outcome tables. Brackets in paired tables apply the same transformation to 95% Student-t intervals on log ratios, or to CT's descriptive fold range. Seed intervals condition on the fixed split. Synthetic intervals vary generated problems, and expansion intervals condition on the shared source data. RQ2 intervals also condition on its fixed tuning split and selected settings. None incorporates uncertainty from new test samples.

Aggregate figures give equal weight to each experiment after within-experiment geometric averaging. RQ2 resamples its 14 families, using 10,000 shared resamples across all settings and both outcomes (PCG64 seed 20260919), with pointwise 2.5th and 97.5th percentiles. These ranges describe sensitivity to benchmark composition conditional on the measured effects, not a joint confidence region, a simultaneous band, or a significance test.

Implementation and evaluation details. LightGBM-based models share native preprocessing and EFB, depth 6, at most 64 leaves, maximum bin count 255, minimum leaf size 20, L2 penalty 1, and GOSS rates 0.2/0.1 (Ke et al., 2017). XGBoost uses CPU histogram training with depth 6, and CatBoost uses depth-6 symmetric trees. Their different samplers do not imply identical realized feature counts. Fits run serially on fixed physical cores of an Intel Xeon Gold 6242 server: eight cores for RQ1 and sixteen for RQ2, identically for all models within each arm. Resource checks accept or retry each fit separately. RQ1 fixes learning rate 0.05 and $n_{\mathrm{ref}} = 300$, with $\delta = 0.05$. Identical lightweight 25-round timing markers are included within each arm's training measurements, without alternative-mask diagnostics. Untimed resource admission precedes each fit. A two-second monitor checks residual CPU activity, I/O and memory pressure, swap activity, and CPU affinity. Only the first accepted fit is retained. Final-test prediction and scoring use saved models after training is complete and do not enter the training-time comparison.

J.1 RQ1: Endogenous and Matched-Control Results

One untimed Endogenous CRAFT-GBM pre-run per training set supplies its round count $t^{(0)}$ and average feature fraction $y^{(0)} = (pt^{(0)})^{-1} \sum_{s=1}^{t^{(0)}} K_s$. LightGBM-Full is the paired reference, using all features for $t^{(0)}$ rounds. LightGBM-Fixed uses this endpoint $t^{(0)}$ and an integer feature count chosen to minimize cumulative feature-count mismatch. XGBoost and CatBoost receive the same round count $t^{(0)}$ and fraction $y^{(0)}$ through native sampling parameters, without exact-count matching. The measured Endogenous CRAFT-GBM fit selects and stops independently, without receiving the pre-run's fraction $y^{(0)}$ or horizon $t^{(0)}$. Both endogenous pre-runs and measured fits use 20 full-feature rounds, if reached. The training-loss stopping rule is unchanged.

Table 3: RQ1: aggregate paired percentage changes. The 18 experiments receive equal weight and belong to 14 source families, and this aggregate is descriptive.

Model	Δ time (%)	Δ loss (%)
Endogenous CRAFT-GBM	-39.74	-1.04
LightGBM-Fixed	-30.91	-1.18
XGBoost	148.74	4.57
CatBoost	153.23	5.56

Table 4: RQ1: frozen pre-run feature budgets and horizons. Values are means [minimum, maximum] across the five paired blocks per experiment. Fixed-split public datasets share one pre-run across model seeds, while other blocks use their own pre-run.

Dataset	p	Pre-run rounds $t^{(0)}$	Pre-run fraction $y^{(0)}$ (%)
HIGGS	28	188.0	59.65
Epsilon	2000	1128.0	16.40
Bosch	968	1339.0	31.64
YearMSD	90	270.0	69.77
APS Failure	170	364.0	59.90
KDD 2009	212	517.0	22.89
Gisette	4326	257.0	39.73
Jannis	54	1612.0	74.42
HAR	561	111.0	94.21
CT Scan	378	343.8 [289.0, 400.0]	50.69 [47.16, 55.50]
Sparse 1	2000	932.2 [910.0, 955.0]	18.80 [18.62, 19.11]
Sparse 2	2000	615.6 [576.0, 656.0]	21.42 [21.01, 21.93]
Dense 1	2000	171.2 [164.0, 181.0]	43.83 [42.07, 45.12]
Dense 2	2000	1108.2 [967.0, 1192.0]	19.48 [17.31, 20.97]
HAR +4	2805	111.2 [110.0, 112.0]	72.26 [71.71, 72.89]
HAR_N	5049	111.6 [110.0, 116.0]	71.74 [69.56, 72.68]
YearMSD +4	450	326.8 [299.0, 346.0]	37.80 [36.61, 39.33]
YearMSD_N	810	316.8 [302.0, 326.0]	32.99 [32.22, 33.85]

Table 5: RQ1: absolute outcomes for every reported configuration. Each row summarizes five replicates: mean $\pm$ sample SD, except CT, which uses mean [minimum, maximum].

Model	Mean rounds	Training (s)	Test loss
HIGGS — Binary log loss (nats/observation)			
LightGBM-Full	188.0	52.91 ± 0.13	$0.5268 \pm 1.89 \times 10^{-4}$
Endogenous CRAFT-GBM	199.4	51.03 ± 1.49	$0.5288 \pm 9.72 \times 10^{-4}$
LightGBM-Fixed	188.0	49.16 ± 0.34	$0.5282 \pm 2.00 \times 10^{-4}$
XGBoost	188.0	75.37 ± 0.29	$0.5288 \pm 5.08 \times 10^{-4}$
CatBoost	188.0	78.57 ± 0.21	$0.5449 \pm 4.22 \times 10^{-5}$
Epsilon — Binary log loss (nats/observation)			
LightGBM-Full	1128.0	221.55 ± 0.24	$0.2849 \pm 1.96 \times 10^{-4}$
Endogenous CRAFT-GBM	1139.6	54.58 ± 0.62	$0.2804 \pm 6.29 \times 10^{-4}$
LightGBM-Fixed	1128.0	72.21 ± 0.19	$0.2886 \pm 1.78 \times 10^{-4}$
XGBoost	1128.0	786.48 ± 1.42	$0.2904 \pm 3.52 \times 10^{-4}$
CatBoost	1128.0	198.32 ± 0.48	$0.3020 \pm 1.05 \times 10^{-4}$
Bosch — Binary log loss (nats/observation)			
LightGBM-Full	1339.0	196.22 ± 0.20	$0.0313 \pm 4.04 \times 10^{-5}$
Endogenous CRAFT-GBM	1470.8	99.84 ± 15.18	$0.0314 \pm 8.93 \times 10^{-5}$
LightGBM-Fixed	1339.0	176.08 ± 0.23	$0.0314 \pm 3.04 \times 10^{-5}$
XGBoost	1339.0	105.97 ± 0.69	$0.0314 \pm 2.36 \times 10^{-5}$
CatBoost	1339.0	541.07 ± 0.99	$0.0317 \pm 1.68 \times 10^{-5}$

Continued on next page

Table 5 (continued)

Model	Mean rounds	Training (s)	Test loss
YearMSD — RMSE (target units)			
LightGBM-Full	270.0	5.36 ± 0.02	9.0259 ± 0.0022
Endogenous CRAFT-GBM	252.4	4.53 ± 0.21	9.0630 ± 0.0062
LightGBM-Fixed	270.0	4.70 ± 0.02	9.0191 ± 0.0037
XGBoost	270.0	8.96 ± 0.15	9.0255 ± 0.0029
CatBoost	270.0	9.48 ± 0.01	9.1688 ± 0.0035
APS Failure — Binary log loss (nats/observation)			
LightGBM-Full	364.0	3.21 ± 0.02	$0.0210 \pm 2.13 \times 10^{-4}$
Endogenous CRAFT-GBM	347.2	2.52 ± 0.04	$0.0203 \pm 1.02 \times 10^{-4}$
LightGBM-Fixed	364.0	2.87 ± 0.02	$0.0208 \pm 3.20 \times 10^{-4}$
XGBoost	364.0	4.77 ± 0.04	$0.0204 \pm 4.15 \times 10^{-4}$
CatBoost	364.0	6.30 ± 0.11	$0.0199 \pm 2.45 \times 10^{-4}$
KDD 2009 — Binary log loss (nats/observation)			
LightGBM-Full	517.0	3.41 ± 0.02	$0.1164 \pm 6.01 \times 10^{-4}$
Endogenous CRAFT-GBM	536.0	2.31 ± 0.03	$0.1146 \pm 9.77 \times 10^{-4}$
LightGBM-Fixed	517.0	2.47 ± 0.02	$0.0999 \pm 6.80 \times 10^{-4}$
XGBoost	517.0	11.64 ± 0.21	$0.1583 \pm 4.65 \times 10^{-4}$
CatBoost	517.0	15.37 ± 0.24	$0.0726 \pm 3.22 \times 10^{-4}$
Gisette — Binary log loss (nats/observation)			
LightGBM-Full	257.0	12.51 ± 0.02	0.0715 ± 0.0015
Endogenous CRAFT-GBM	262.2	7.24 ± 0.09	0.0700 ± 0.0022
LightGBM-Fixed	257.0	7.56 ± 0.04	$0.0646 \pm 6.56 \times 10^{-4}$
XGBoost	257.0	24.06 ± 0.04	0.0651 ± 0.0016
CatBoost	257.0	36.69 ± 0.33	0.0809 ± 0.0018
Jannis — Multiclass log loss (nats/observation)			
LightGBM-Full	1612.0	15.31 ± 0.05	0.7677 ± 0.0016
Endogenous CRAFT-GBM	1559.2	13.34 ± 0.65	0.7446 ± 0.0064
LightGBM-Fixed	1612.0	14.21 ± 0.12	0.7633 ± 0.0018
XGBoost	1612.0	25.59 ± 0.10	$0.7042 \pm 7.02 \times 10^{-4}$
CatBoost	1612.0	25.83 ± 0.09	$0.6801 \pm 3.98 \times 10^{-4}$
HAR — Multiclass log loss (nats/observation)			
LightGBM-Full	111.0	2.67 ± 0.01	0.1687 ± 0.0036
Endogenous CRAFT-GBM	114.6	2.67 ± 0.04	0.1694 ± 0.0017
LightGBM-Fixed	111.0	2.60 ± 0.00	0.1667 ± 0.0030
XGBoost	111.0	9.62 ± 0.02	0.2181 ± 0.0016
CatBoost	111.0	19.38 ± 0.06	0.2617 ± 0.0029
CT Scan — RMSE (target units)			
LightGBM-Full	343.8	$5.10\ [4.40, 5.72]$	$6.1844\ [4.6835, 7.2040]$
Endogenous CRAFT-GBM	319.2	$3.46\ [3.26, 3.71]$	$6.2439\ [4.8497, 7.1849]$
LightGBM-Fixed	343.8	$3.99\ [3.58, 4.36]$	$6.0185\ [4.4152, 6.9385]$
XGBoost	343.8	$7.05\ [6.66, 7.87]$	$5.9382\ [4.5108, 6.8590]$
CatBoost	343.8	$8.01\ [7.06, 9.18]$	$6.4751\ [5.1111, 7.2789]$
Sparse 1 — Binary log loss (nats/observation)			
LightGBM-Full	932.2	56.75 ± 0.97	0.5041 ± 0.0012
Endogenous CRAFT-GBM	913.2	16.58 ± 0.22	0.4991 ± 0.0013
LightGBM-Fixed	932.2	36.32 ± 0.52	0.5098 ± 0.0015
XGBoost	932.2	108.63 ± 2.28	0.5262 ± 0.0014
CatBoost	932.2	96.43 ± 0.99	0.5262 ± 0.0016

Continued on next page

Table 5 (continued)

Model	Mean rounds	Training (s)	Test loss
Sparse 2 — Binary log loss (nats/observation)			
LightGBM-Full	615.6	2.97 ± 0.11	0.4686 ± 0.0016
Endogenous CRAFT-GBM	579.6	2.57 ± 0.07	0.4650 ± 0.0016
LightGBM-Fixed	615.6	2.89 ± 0.10	0.4738 ± 0.0021
XGBoost	615.6	25.84 ± 0.52	0.4748 ± 0.0023
CatBoost	615.6	20.01 ± 0.55	0.4703 ± 0.0021
Dense 1 — RMSE (target units)			
LightGBM-Full	171.2	15.71 ± 0.59	1.0659 ± 0.0029
Endogenous CRAFT-GBM	176.2	9.11 ± 0.08	1.0565 ± 0.0039
LightGBM-Fixed	171.2	8.97 ± 0.14	1.1664 ± 0.0316
XGBoost	171.2	49.87 ± 1.63	1.1620 ± 0.0212
CatBoost	171.2	28.10 ± 0.73	1.0913 ± 0.0101
Dense 2 — Binary log loss (nats/observation)			
LightGBM-Full	1108.2	78.67 ± 6.67	0.3860 ± 0.0190
Endogenous CRAFT-GBM	1170.8	19.68 ± 1.33	0.3691 ± 0.0205
LightGBM-Fixed	1108.2	24.01 ± 2.32	0.3933 ± 0.0185
XGBoost	1108.2	199.59 ± 13.30	0.3684 ± 0.0169
CatBoost	1108.2	93.92 ± 10.27	0.3733 ± 0.0172
HAR +4 — Multiclass log loss (nats/observation)			
LightGBM-Full	111.2	9.63 ± 0.08	0.1727 ± 0.0039
Endogenous CRAFT-GBM	113.8	7.42 ± 0.07	0.1710 ± 0.0010
LightGBM-Fixed	111.2	7.56 ± 0.03	0.1665 ± 0.0034
XGBoost	111.2	47.96 ± 1.25	0.2082 ± 0.0042
CatBoost	111.2	92.62 ± 1.05	0.2599 ± 0.0028
HAR_N — Multiclass log loss (nats/observation)			
LightGBM-Full	111.6	16.51 ± 0.33	0.1734 ± 0.0031
Endogenous CRAFT-GBM	112.0	12.38 ± 0.21	0.1714 ± 0.0034
LightGBM-Fixed	111.6	12.85 ± 0.05	0.1708 ± 0.0049
XGBoost	111.6	87.46 ± 1.15	0.2085 ± 0.0055
CatBoost	111.6	179.86 ± 3.19	0.2638 ± 0.0041
YearMSD +4 — RMSE (target units)			
LightGBM-Full	326.8	19.70 ± 1.06	9.0245 ± 0.0059
Endogenous CRAFT-GBM	322.0	10.95 ± 0.47	9.0093 ± 0.0050
LightGBM-Fixed	326.8	11.27 ± 0.26	9.0248 ± 0.0041
XGBoost	326.8	57.61 ± 2.44	9.0086 ± 0.0078
CatBoost	326.8	29.92 ± 1.15	9.1362 ± 0.0109
YearMSD_N — RMSE (target units)			
LightGBM-Full	316.8	32.38 ± 1.13	9.0337 ± 0.0050
Endogenous CRAFT-GBM	321.8	16.38 ± 1.07	9.0129 ± 0.0085
LightGBM-Fixed	316.8	16.65 ± 1.01	9.0421 ± 0.0052
XGBoost	316.8	111.67 ± 2.20	9.0222 ± 0.0059
CatBoost	316.8	47.22 ± 0.88	9.1438 ± 0.0069

Table 6: RQ1: paired percentage changes relative to LightGBM-Full. Entries apply $100(r-1)$ to geometric-mean ratios and pointwise paired 95% interval endpoints, whereas CT brackets are descriptive fold ranges. Negative values are favorable.

Model	Δ time (%)	Interval	Δ loss (%)	Interval
HIGGS				
Endogenous CRAFT-GBM	-3.59	[-7.27, 0.23]	0.39	[0.14, 0.64]
LightGBM-Fixed	-7.09	[-8.05, -6.11]	0.28	[0.20, 0.35]
XGBoost	42.44	[41.36, 43.53]	0.39	[0.27, 0.51]
CatBoost	48.49	[48.01, 48.97]	3.44	[3.39, 3.50]
Epsilon				
Endogenous CRAFT-GBM	-75.37	[-75.72, -75.01]	-1.58	[-1.81, -1.35]
LightGBM-Fixed	-67.41	[-67.49, -67.33]	1.31	[1.16, 1.46]
XGBoost	255.00	[254.45, 255.55]	1.93	[1.72, 2.15]
CatBoost	-10.49	[-10.73, -10.24]	6.00	[5.87, 6.12]
Bosch				
Endogenous CRAFT-GBM	-49.61	[-58.50, -38.81]	0.05	[-0.40, 0.51]
LightGBM-Fixed	-10.26	[-10.32, -10.20]	0.14	[-0.07, 0.35]
XGBoost	-45.99	[-46.38, -45.61]	0.08	[-0.08, 0.25]
CatBoost	175.75	[175.17, 176.34]	1.04	[0.89, 1.19]
YearMSD				
Endogenous CRAFT-GBM	-15.55	[-20.44, -10.36]	0.41	[0.35, 0.47]
LightGBM-Fixed	-12.28	[-12.52, -12.04]	-0.08	[-0.13, -0.02]
XGBoost	67.20	[64.10, 70.37]	0.00	[-0.07, 0.06]
CatBoost	76.86	[76.36, 77.35]	1.58	[1.52, 1.65]
APS Failure				
Endogenous CRAFT-GBM	-21.48	[-22.91, -20.02]	-3.20	[-4.85, -1.53]
LightGBM-Fixed	-10.71	[-11.80, -9.60]	-0.90	[-3.06, 1.31]
XGBoost	48.51	[46.12, 50.93]	-2.65	[-6.00, 0.82]
CatBoost	96.18	[91.19, 101.31]	-5.13	[-6.46, -3.78]
KDD 2009				
Endogenous CRAFT-GBM	-32.19	[-33.26, -31.10]	-1.58	[-3.04, -0.10]
LightGBM-Fixed	-27.45	[-28.35, -26.55]	-14.23	[-14.88, -13.58]
XGBoost	241.64	[233.78, 249.68]	35.91	[35.04, 36.78]
CatBoost	351.08	[341.58, 360.77]	-37.67	[-38.21, -37.14]
Gisette				
Endogenous CRAFT-GBM	-42.17	[-43.17, -41.15]	-2.05	[-4.50, 0.45]
LightGBM-Fixed	-39.59	[-40.01, -39.16]	-9.66	[-12.49, -6.74]
XGBoost	92.29	[91.94, 92.64]	-8.99	[-11.63, -6.26]
CatBoost	193.21	[190.08, 196.38]	13.13	[8.51, 17.94]
Jannis				
Endogenous CRAFT-GBM	-12.96	[-18.02, -7.57]	-3.02	[-4.15, -1.87]
LightGBM-Fixed	-7.21	[-8.19, -6.21]	-0.58	[-1.06, -0.10]
XGBoost	67.12	[66.50, 67.75]	-8.27	[-8.57, -7.97]
CatBoost	68.66	[68.09, 69.22]	-11.41	[-11.70, -11.12]
HAR				
Endogenous CRAFT-GBM	0.11	[-1.83, 2.09]	0.44	[-2.35, 3.32]
LightGBM-Fixed	-2.72	[-2.96, -2.48]	-1.19	[-4.92, 2.69]
XGBoost	260.01	[258.58, 261.44]	29.33	[26.62, 32.09]
CatBoost	625.35	[622.02, 628.69]	55.16	[50.84, 59.60]

Continued on next page

Table 6 (continued)

Model	Δ time (%)	Interval	Δ loss (%)	Interval
CT Scan — descriptive fold ranges				
Endogenous CRAFT-GBM	-31.99	[-40.37, -25.90]	1.09	[-0.26, 3.55]
LightGBM-Fixed	-21.68	[-23.74, -18.62]	-2.73	[-9.77, 1.16]
XGBoost	38.56	[28.77, 51.24]	-3.88	[-7.60, -0.63]
CatBoost	57.10	[48.12, 60.59]	5.07	[1.04, 9.13]
Sparse 1				
Endogenous CRAFT-GBM	-70.78	[-71.80, -69.72]	-1.01	[-1.40, -0.61]
LightGBM-Fixed	-36.00	[-36.23, -35.76]	1.12	[0.97, 1.26]
XGBoost	91.40	[87.51, 95.37]	4.38	[4.25, 4.51]
CatBoost	69.93	[68.38, 71.49]	4.38	[4.24, 4.51]
Sparse 2				
Endogenous CRAFT-GBM	-13.64	[-14.95, -12.31]	-0.78	[-0.85, -0.72]
LightGBM-Fixed	-2.80	[-3.93, -1.65]	1.11	[0.94, 1.27]
XGBoost	769.83	[735.33, 805.75]	1.30	[1.08, 1.53]
CatBoost	573.55	[562.27, 585.03]	0.36	[0.16, 0.56]
Dense 1				
Endogenous CRAFT-GBM	-41.96	[-44.36, -39.46]	-0.87	[-1.29, -0.46]
LightGBM-Fixed	-42.84	[-44.46, -41.17]	9.40	[5.77, 13.16]
XGBoost	217.58	[213.57, 221.65]	9.00	[6.78, 11.27]
CatBoost	78.99	[76.16, 81.87]	2.38	[1.37, 3.41]
Dense 2				
Endogenous CRAFT-GBM	-74.95	[-77.85, -71.68]	-4.41	[-5.62, -3.18]
LightGBM-Fixed	-69.50	[-70.79, -68.16]	1.91	[0.91, 2.92]
XGBoost	153.99	[147.68, 160.46]	-4.55	[-5.08, -4.02]
CatBoost	19.16	[11.82, 26.99]	-3.28	[-4.19, -2.36]
HAR +4				
Endogenous CRAFT-GBM	-22.96	[-24.34, -21.55]	-0.96	[-3.24, 1.38]
LightGBM-Fixed	-21.46	[-22.19, -20.73]	-3.56	[-8.20, 1.32]
XGBoost	398.04	[378.67, 418.20]	20.59	[16.61, 24.72]
CatBoost	861.97	[854.57, 869.43]	50.53	[45.02, 56.25]
HAR_N				
Endogenous CRAFT-GBM	-24.99	[-27.44, -22.47]	-1.15	[-2.32, 0.03]
LightGBM-Fixed	-22.14	[-23.71, -20.55]	-1.54	[-4.59, 1.61]
XGBoost	429.84	[409.06, 451.47]	20.24	[15.99, 24.65]
CatBoost	989.54	[980.86, 998.29]	52.13	[49.53, 54.77]
YearMSD +4				
Endogenous CRAFT-GBM	-44.39	[-48.91, -39.48]	-0.17	[-0.24, -0.10]
LightGBM-Fixed	-42.70	[-47.72, -37.19]	0.00	[-0.06, 0.07]
XGBoost	192.66	[177.38, 208.78]	-0.18	[-0.26, -0.09]
CatBoost	51.99	[46.87, 57.29]	1.24	[1.13, 1.34]
YearMSD_N				
Endogenous CRAFT-GBM	-49.47	[-54.55, -43.81]	-0.23	[-0.31, -0.15]
LightGBM-Fixed	-48.62	[-53.61, -43.10]	0.09	[0.02, 0.17]
XGBoost	245.01	[230.90, 259.73]	-0.13	[-0.20, -0.05]
CatBoost	45.89	[40.66, 51.32]	1.22	[1.13, 1.30]

Table 7: RQ1: synthetic oracle diagnostics used only for evaluation. Values are mean ± sample SD over five generated problems. The first metric is oracle cross-entropy for Sparse 1/2 and function MSE for Dense 1, while Bernoulli KL is defined only for Sparse 1/2. These quantities do not guide training or selection.

Model	Oracle CE / function MSE	Bernoulli KL
Sparse 1 — Oracle cross-entropy and Bernoulli KL (nats/observation)		
LightGBM-Full	0.5035 ± 0.0013	$0.0312 \pm 5.50\times10^{-4}$
Endogenous CRAFT-GBM	0.4984 ± 0.0018	0.0260 ± 0.0013
LightGBM-Fixed	0.5094 ± 0.0014	$0.0370 \pm 7.04\times10^{-4}$
XGBoost	0.5259 ± 0.0016	$0.0535 \pm 9.64\times10^{-4}$
CatBoost	0.5258 ± 0.0016	0.0534 ± 0.0011
Sparse 2 — Oracle cross-entropy and Bernoulli KL (nats/observation)		
LightGBM-Full	$0.4681 \pm 8.17\times10^{-4}$	$0.0135 \pm 9.08\times10^{-4}$
Endogenous CRAFT-GBM	$0.4646 \pm 7.04\times10^{-4}$	$0.0099 \pm 7.55\times10^{-4}$
LightGBM-Fixed	0.4733 ± 0.0013	0.0187 ± 0.0016
XGBoost	0.4744 ± 0.0016	0.0197 ± 0.0018
CatBoost	0.4699 ± 0.0015	0.0153 ± 0.0017
Dense 1 — Function MSE (squared target units)		
LightGBM-Full	0.1355 ± 0.0050	—
Endogenous CRAFT-GBM	0.1162 ± 0.0078	—
LightGBM-Fixed	0.3633 ± 0.0761	—
XGBoost	0.3520 ± 0.0439	—
CatBoost	0.1918 ± 0.0196	—

J.2 RQ2: Exogenous Fractions and Frozen Tuning Selections

Within each campaign, the same LightGBM-Full reference serves all five fractions f in a paired comparison. Exogenous CRAFT-GBM uses interval point estimates $c_{j,t}$ and its fresh same-campaign Full reference, while the other non-ablated models retain their original paired reference. The original, updated CRAFT-GBM, and ablation Full references are listed separately in the absolute-outcome tables. Data, tuning selections, and training settings are shared across these campaigns. CRAFT-GBM's first two rounds use all features, whereas its later rounds and LightGBM-Fixed use the same target $K_f = \max\{1, \lfloor fp + 1/2 \rfloor\}$, encoded through each implementation's native fraction parameter f. XGBoost/CatBoost instead receive f directly. RQ2 fixes $\delta = 0.05$ and $n_{\text{ref}} = 1000$. LightGBM-Full's round count $T^\star$ and learning rate $\eta^\star$ are selected on an 80/20 inner training/validation split, group-disjoint for HAR and CT. All measured models then refit on the full training set with that selected pair $(T^\star, \eta^\star)$. The 34 tuning studies use 20 trials each: rounds in $[50, 2000]$ and learning rate in $[0.01, 0.20]$, both searched on logarithmic scales. Optuna's seeded TPE search (**??**) includes $(300, 0.05)$, uses five startup trials, and selects by inner-validation loss at the requested endpoint, without early stopping. The inner-split seed is 20260916, the TPE seed is 20260917, and the tuning model seed is 20260918. Measured refits use the measurement seeds above. Tuning uses eight cores and is excluded from measured sixteen-core times. Selected rates $\eta^\star$ below are rounded only for display, while fitting uses full precision.

Table 8: RQ2: aggregate paired percentage changes. The 14 original families receive equal weight. Brackets are pointwise 95% family-resampling bands describing benchmark-composition sensitivity.

Model	f	Δ time (%)	Δ loss (%)
Exogenous CRAFT-GBM	0.1	-62.74 [-73.74, -47.52]	-0.04 [-2.83, 2.70]
Exogenous CRAFT-GBM	0.3	-44.20 [-54.21, -32.29]	-1.16 [-2.76, 0.29]
Exogenous CRAFT-GBM	0.5	-28.02 [-36.05, -19.37]	-0.27 [-1.77, 1.71]

Continued on next page

Table 8 (continued)

Model	f	Δ time (%)	Δ loss (%)
Exogenous CRAFT-GBM	0.7	-11.77 [-17.23, -6.25]	-0.38 [-1.48, 0.96]
Exogenous CRAFT-GBM	0.9	-1.81 [-4.02, 0.51]	-0.26 [-0.49, -0.05]
LightGBM-Fixed	0.1	-49.22 [-60.91, -35.28]	4.60 [-3.04, 17.28]
LightGBM-Fixed	0.3	-35.17 [-44.82, -24.58]	-1.44 [-4.11, 1.05]
LightGBM-Fixed	0.5	-23.25 [-30.28, -15.96]	-1.34 [-2.76, -0.12]
LightGBM-Fixed	0.7	-9.41 [-14.16, -5.19]	-0.88 [-1.72, -0.20]
LightGBM-Fixed	0.9	-2.79 [-4.37, -1.42]	-0.59 [-1.10, -0.15]
XGBoost	0.1	90.61 [39.57, 166.55]	3.31 [-3.31, 13.25]
XGBoost	0.3	102.74 [49.16, 182.34]	-0.17 [-3.01, 3.11]
XGBoost	0.5	112.23 [55.84, 194.82]	0.28 [-2.09, 3.43]
XGBoost	0.7	122.12 [63.47, 207.78]	0.79 [-1.40, 3.88]
XGBoost	0.9	130.74 [69.92, 218.10]	1.52 [-0.60, 4.48]
CatBoost	0.1	29.20 [-3.88, 82.51]	4.44 [1.86, 8.10]
CatBoost	0.3	120.07 [68.11, 201.68]	2.42 [0.92, 4.28]
CatBoost	0.5	180.62 [114.59, 285.03]	2.20 [0.54, 4.12]
CatBoost	0.7	207.04 [134.29, 321.14]	2.15 [0.49, 4.07]
CatBoost	0.9	201.02 [129.26, 313.24]	2.14 [0.42, 4.15]
Mean ablation	0.1	-63.83 [-74.42, -49.53]	12.69 [0.13, 35.01]
Mean ablation	0.3	-45.08 [-54.84, -33.65]	7.73 [-0.31, 22.75]
Mean ablation	0.5	-29.72 [-37.36, -21.46]	4.40 [-0.43, 13.28]
Mean ablation	0.7	-13.19 [-18.46, -8.08]	3.53 [-0.51, 11.47]
Mean ablation	0.9	-3.63 [-5.59, -1.65]	0.41 [-0.14, 1.20]
Exposure ablation	0.1	-64.71 [-75.48, -49.48]	5.71 [-0.90, 16.49]
Exposure ablation	0.3	-47.00 [-56.87, -34.68]	2.16 [-2.98, 10.67]
Exposure ablation	0.5	-31.73 [-39.55, -22.95]	1.94 [-1.86, 8.39]
Exposure ablation	0.7	-13.82 [-18.94, -8.93]	-0.99 [-2.02, -0.10]
Exposure ablation	0.9	-4.38 [-6.47, -2.39]	-0.41 [-0.81, -0.08]

Table 9: RQ2: all 34 frozen tuning selections. Each study has 20 valid, unique trials (680 total). $T^\star$ and $\eta^\star$ are transferred to every measured method and fraction. Boundary marks a selected search limit, while tuning wall time is separate from measured training time.

Dataset / identity	$T^\star$	$\eta^\star$	Validation loss	Boundary	Tuning (s)
HIGGS	1955	0.152296	0.48986	none	4504.9
Epsilon	1842	0.066851	0.27077	none	3556.7
Bosch	1780	0.013433	0.03087	none	2448.3
YearMSD	1865	0.065079	8.72367	none	436.4
APS Failure	571	0.023038	0.01752	none	146.5
KDD 2009	93	0.022992	0.07730	none	116.2
Gisette	1652	0.010419	0.06000	none	668.9
Jannis	862	0.013581	0.70006	none	348.4
HAR	449	0.101535	0.14463	none	411.6
CT Scan / fold 01	1319	0.032177	7.61767	none	270.0
CT Scan / fold 02	1122	0.026032	4.82242	none	256.9
CT Scan / fold 03	666	0.194719	8.74659	none	236.0
CT Scan / fold 04	857	0.029059	4.23218	none	266.4
CT Scan / fold 05	1704	0.021218	7.72786	none	273.9
Sparse 1 / rep. 00	1625	0.037491	0.50314	none	1150.8
Sparse 1 / rep. 01	1246	0.056201	0.50218	none	918.6
Sparse 1 / rep. 02	1940	0.034574	0.50015	none	1173.7
Sparse 1 / rep. 03	1940	0.034574	0.49865	none	1211.6
Sparse 1 / rep. 04	938	0.084292	0.49886	none	960.7

Continued on next page

Table 9 (continued)

Dataset / identity	$T^\star$	$\eta^\star$	Validation loss	Boundary	Tuning (s)
Sparse 2 / rep. 00	1894	0.022050	0.46350	none	173.5
Sparse 2 / rep. 01	577	0.053795	0.46514	none	158.6
Sparse 2 / rep. 02	1705	0.023268	0.46185	none	178.7
Sparse 2 / rep. 03	816	0.047642	0.46192	none	160.0
Sparse 2 / rep. 04	1705	0.023268	0.46700	none	186.0
Dense 1 / rep. 00	1787	0.013074	1.04110	none	1206.0
Dense 1 / rep. 01	1667	0.010415	1.04943	none	1193.2
Dense 1 / rep. 02	1667	0.010415	1.05503	none	1213.0
Dense 1 / rep. 03	1780	0.013433	1.05663	none	1427.5
Dense 1 / rep. 04	1395	0.012046	1.05622	none	1214.3
Dense 2 / rep. 00	1944	0.012398	0.37479	none	1237.1
Dense 2 / rep. 01	1780	0.013433	0.38733	none	1322.5
Dense 2 / rep. 02	1780	0.013433	0.36559	none	2538.6
Dense 2 / rep. 03	1780	0.013433	0.36908	none	2466.3
Dense 2 / rep. 04	1780	0.013433	0.41292	none	1838.3

Table 10: RQ2: absolute outcomes for every measured configuration. Each row summarizes five replicates: mean ± sample SD, except CT, which uses mean [minimum, maximum]. Full (CRAFT-GBM ref.) is the reference for Exogenous CRAFT-GBM, Full (ablation ref.) serves Mean and Exposure ablations, and the unqualified Full row serves the remaining models. Each reference was measured in its own campaign.

Model	f	Mean rounds	Training (s)	Test loss
HIGGS — Binary log loss (nats/observation)				
LightGBM-Full	—	1955.0	282.24 ± 0.48	$0.4873 \pm 1.45 \times 10^{-4}$
LightGBM-Full (CRAFT-GBM ref.)	—	1955.0	281.97 ± 0.50	$0.4873 \pm 1.45 \times 10^{-4}$
LightGBM-Full (ablation ref.)	—	1955.0	281.05 ± 0.28	$0.4873 \pm 1.45 \times 10^{-4}$
Exogenous CRAFT-GBM	0.1	1955.0	219.95 ± 1.48	0.5343 ± 0.0028
Exogenous CRAFT-GBM	0.3	1955.0	232.85 ± 0.43	0.5071 ± 0.0022
Exogenous CRAFT-GBM	0.5	1955.0	243.14 ± 1.09	$0.5000 \pm 3.32 \times 10^{-4}$
Exogenous CRAFT-GBM	0.7	1955.0	266.25 ± 0.88	$0.4903 \pm 4.46 \times 10^{-4}$
Exogenous CRAFT-GBM	0.9	1955.0	273.98 ± 0.36	$0.4897 \pm 5.15 \times 10^{-4}$
LightGBM-Fixed	0.1	1955.0	240.45 ± 0.99	$0.5156 \pm 4.35 \times 10^{-4}$
LightGBM-Fixed	0.3	1955.0	249.26 ± 1.71	$0.4965 \pm 8.85 \times 10^{-4}$
LightGBM-Fixed	0.5	1955.0	262.32 ± 0.38	$0.4908 \pm 1.93 \times 10^{-4}$
LightGBM-Fixed	0.7	1955.0	274.86 ± 0.79	$0.4888 \pm 2.37 \times 10^{-4}$
LightGBM-Fixed	0.9	1955.0	278.93 ± 0.49	$0.4879 \pm 3.15 \times 10^{-4}$
XGBoost	0.1	1955.0	381.16 ± 0.91	$0.5276 \pm 9.27 \times 10^{-4}$
XGBoost	0.3	1955.0	389.57 ± 1.55	$0.4906 \pm 2.13 \times 10^{-4}$
XGBoost	0.5	1955.0	385.12 ± 1.13	$0.4845 \pm 1.81 \times 10^{-4}$
XGBoost	0.7	1955.0	385.01 ± 3.75	$0.4826 \pm 1.75 \times 10^{-4}$
XGBoost	0.9	1955.0	384.26 ± 3.75	$0.4812 \pm 3.00 \times 10^{-4}$
CatBoost	0.1	1955.0	376.20 ± 2.17	$0.5038 \pm 2.75 \times 10^{-4}$
CatBoost	0.3	1955.0	467.00 ± 1.39	$0.4936 \pm 1.33 \times 10^{-4}$
CatBoost	0.5	1955.0	524.75 ± 2.76	$0.4918 \pm 9.23 \times 10^{-5}$
CatBoost	0.7	1955.0	552.83 ± 3.14	$0.4910 \pm 9.70 \times 10^{-5}$
CatBoost	0.9	1955.0	555.91 ± 2.84	$0.4907 \pm 5.89 \times 10^{-5}$
Mean ablation	0.1	1955.0	218.63 ± 0.37	0.5335 ± 0.0029
Mean ablation	0.3	1955.0	228.80 ± 1.33	0.5254 ± 0.0020
Mean ablation	0.5	1955.0	254.36 ± 2.18	$0.5019 \pm 1.41 \times 10^{-4}$
Mean ablation	0.7	1955.0	271.31 ± 0.32	$0.4948 \pm 1.24 \times 10^{-4}$
Mean ablation	0.9	1955.0	273.38 ± 0.73	$0.4887 \pm 4.67 \times 10^{-5}$
Exposure ablation	0.1	1955.0	225.90 ± 1.02	$0.5307 \pm 1.75 \times 10^{-5}$
Exposure ablation	0.3	1955.0	237.17 ± 0.42	$0.5058 \pm 6.96 \times 10^{-5}$
Exposure ablation	0.5	1955.0	250.64 ± 0.46	$0.5025 \pm 6.34 \times 10^{-5}$
Exposure ablation	0.7	1955.0	269.68 ± 0.68	$0.4880 \pm 2.20 \times 10^{-4}$
Exposure ablation	0.9	1955.0	275.62 ± 1.47	$0.4877 \pm 1.63 \times 10^{-4}$

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
Epsilon — Binary log loss (nats/observation)				
LightGBM-Full	—	1842.0	214.78 ± 0.36	$0.2738 \pm 3.72 \times 10^{-4}$
LightGBM-Full (CRAFT-GBM ref.)	—	1842.0	214.58 ± 0.58	$0.2738 \pm 3.72 \times 10^{-4}$
LightGBM-Full (ablation ref.)	—	1842.0	213.96 ± 0.29	$0.2738 \pm 3.72 \times 10^{-4}$
Exogenous CRAFT-GBM	0.1	1842.0	34.01 ± 0.16	$0.2674 \pm 3.97 \times 10^{-4}$
Exogenous CRAFT-GBM	0.3	1842.0	75.84 ± 0.19	$0.2699 \pm 7.19 \times 10^{-4}$
Exogenous CRAFT-GBM	0.5	1842.0	115.73 ± 0.32	$0.2710 \pm 7.16 \times 10^{-4}$
Exogenous CRAFT-GBM	0.7	1842.0	155.82 ± 0.48	$0.2720 \pm 4.92 \times 10^{-4}$
Exogenous CRAFT-GBM	0.9	1842.0	196.85 ± 0.29	$0.2728 \pm 4.49 \times 10^{-4}$
LightGBM-Fixed	0.1	1842.0	58.05 ± 0.07	$0.2754 \pm 3.76 \times 10^{-4}$
LightGBM-Fixed	0.3	1842.0	94.84 ± 0.14	$0.2731 \pm 3.87 \times 10^{-4}$
LightGBM-Fixed	0.5	1842.0	130.22 ± 0.41	$0.2732 \pm 5.28 \times 10^{-4}$
LightGBM-Fixed	0.7	1842.0	164.79 ± 0.26	$0.2731 \pm 1.74 \times 10^{-4}$
LightGBM-Fixed	0.9	1842.0	199.10 ± 0.26	$0.2731 \pm 3.50 \times 10^{-4}$
XGBoost	0.1	1842.0	770.94 ± 1.85	$0.2718 \pm 1.86 \times 10^{-4}$
XGBoost	0.3	1842.0	772.47 ± 2.82	$0.2720 \pm 4.20 \times 10^{-4}$
XGBoost	0.5	1842.0	784.94 ± 1.40	$0.2727 \pm 1.14 \times 10^{-4}$
XGBoost	0.7	1842.0	800.80 ± 1.29	$0.2723 \pm 3.32 \times 10^{-4}$
XGBoost	0.9	1842.0	817.53 ± 2.23	$0.2727 \pm 2.45 \times 10^{-4}$
CatBoost	0.1	1842.0	148.04 ± 0.34	$0.2760 \pm 9.85 \times 10^{-5}$
CatBoost	0.3	1842.0	331.78 ± 0.41	$0.2754 \pm 2.02 \times 10^{-4}$
CatBoost	0.5	1842.0	456.68 ± 0.69	$0.2753 \pm 1.63 \times 10^{-4}$
CatBoost	0.7	1842.0	515.14 ± 0.71	$0.2753 \pm 1.86 \times 10^{-4}$
CatBoost	0.9	1842.0	516.03 ± 0.53	$0.2752 \pm 2.14 \times 10^{-4}$
Mean ablation	0.1	1842.0	32.27 ± 0.06	$0.2766 \pm 5.39 \times 10^{-4}$
Mean ablation	0.3	1842.0	74.29 ± 0.18	$0.2702 \pm 3.69 \times 10^{-4}$
Mean ablation	0.5	1842.0	114.14 ± 0.18	$0.2715 \pm 5.22 \times 10^{-4}$
Mean ablation	0.7	1842.0	154.71 ± 0.27	$0.2723 \pm 6.01 \times 10^{-4}$
Mean ablation	0.9	1842.0	195.03 ± 0.50	$0.2729 \pm 5.10 \times 10^{-4}$
Exposure ablation	0.1	1842.0	33.10 ± 0.07	$0.2786 \pm 4.68 \times 10^{-4}$
Exposure ablation	0.3	1842.0	74.50 ± 0.08	$0.2738 \pm 2.43 \times 10^{-4}$
Exposure ablation	0.5	1842.0	114.81 ± 0.33	$0.2736 \pm 3.30 \times 10^{-4}$
Exposure ablation	0.7	1842.0	154.74 ± 0.17	$0.2731 \pm 1.88 \times 10^{-4}$
Exposure ablation	0.9	1842.0	194.50 ± 0.32	$0.2731 \pm 5.48 \times 10^{-4}$

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
Bosch — Binary log loss (nats/observation)				
LightGBM-Full	—	1780.0	154.08 ± 0.55	$0.0312 \pm 1.33 \times 10^{-5}$
LightGBM-Full (CRAFT-GBM ref.)	—	1780.0	153.89 ± 0.38	$0.0312 \pm 1.33 \times 10^{-5}$
LightGBM-Full (ablation ref.)	—	1780.0	153.46 ± 0.18	$0.0312 \pm 1.33 \times 10^{-5}$
Exogenous CRAFT-GBM	0.1	1780.0	44.16 ± 0.37	$0.0315 \pm 7.26 \times 10^{-6}$
Exogenous CRAFT-GBM	0.3	1780.0	72.46 ± 0.33	$0.0313 \pm 1.45 \times 10^{-5}$
Exogenous CRAFT-GBM	0.5	1780.0	103.84 ± 0.42	$0.0313 \pm 1.27 \times 10^{-5}$
Exogenous CRAFT-GBM	0.7	1780.0	128.68 ± 0.35	$0.0312 \pm 1.99 \times 10^{-5}$
Exogenous CRAFT-GBM	0.9	1780.0	145.51 ± 0.13	$0.0312 \pm 1.11 \times 10^{-5}$
LightGBM-Fixed	0.1	1780.0	117.15 ± 0.12	$0.0314 \pm 9.81 \times 10^{-6}$
LightGBM-Fixed	0.3	1780.0	127.63 ± 0.45	$0.0313 \pm 3.53 \times 10^{-5}$
LightGBM-Fixed	0.5	1780.0	138.13 ± 0.25	$0.0312 \pm 1.02 \times 10^{-5}$
LightGBM-Fixed	0.7	1780.0	146.77 ± 0.49	$0.0312 \pm 1.40 \times 10^{-5}$
LightGBM-Fixed	0.9	1780.0	151.19 ± 0.19	$0.0312 \pm 1.98 \times 10^{-5}$
XGBoost	0.1	1780.0	97.29 ± 0.17	$0.0314 \pm 1.79 \times 10^{-5}$
XGBoost	0.3	1780.0	99.67 ± 0.41	$0.0313 \pm 1.74 \times 10^{-5}$
XGBoost	0.5	1780.0	103.61 ± 0.38	$0.0313 \pm 1.78 \times 10^{-5}$
XGBoost	0.7	1780.0	107.35 ± 0.59	$0.0313 \pm 1.24 \times 10^{-5}$
XGBoost	0.9	1780.0	111.22 ± 0.35	$0.0313 \pm 1.35 \times 10^{-5}$
CatBoost	0.1	1780.0	187.09 ± 0.97	$0.0319 \pm 9.57 \times 10^{-6}$
CatBoost	0.3	1780.0	406.46 ± 1.70	$0.0318 \pm 1.10 \times 10^{-5}$
CatBoost	0.5	1780.0	536.06 ± 0.72	$0.0318 \pm 5.38 \times 10^{-6}$
CatBoost	0.7	1780.0	563.26 ± 1.96	$0.0318 \pm 7.55 \times 10^{-6}$
CatBoost	0.9	1780.0	495.47 ± 1.96	$0.0318 \pm 1.02 \times 10^{-5}$
Mean ablation	0.1	1780.0	43.57 ± 0.20	$0.0316 \pm 8.71 \times 10^{-6}$
Mean ablation	0.3	1780.0	77.52 ± 0.08	$0.0313 \pm 1.49 \times 10^{-5}$
Mean ablation	0.5	1780.0	103.19 ± 0.15	$0.0312 \pm 1.99 \times 10^{-5}$
Mean ablation	0.7	1780.0	127.26 ± 0.14	$0.0312 \pm 2.46 \times 10^{-5}$
Mean ablation	0.9	1780.0	144.10 ± 0.19	$0.0312 \pm 3.09 \times 10^{-5}$
Exposure ablation	0.1	1780.0	39.65 ± 0.09	$0.0315 \pm 9.02 \times 10^{-6}$
Exposure ablation	0.3	1780.0	68.40 ± 0.09	$0.0313 \pm 6.18 \times 10^{-6}$
Exposure ablation	0.5	1780.0	98.77 ± 0.30	$0.0314 \pm 3.30 \times 10^{-6}$
Exposure ablation	0.7	1780.0	123.10 ± 0.06	$0.0312 \pm 9.84 \times 10^{-6}$
Exposure ablation	0.9	1780.0	142.33 ± 0.37	$0.0312 \pm 1.24 \times 10^{-5}$

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
YearMSD — RMSE (target units)				
LightGBM-Full	—	1865.0	16.95 ± 0.04	8.9440 ± 0.0045
LightGBM-Full (CRAFT-GBM ref.)	—	1865.0	16.90 ± 0.10	8.9440 ± 0.0045
LightGBM-Full (ablation ref.)	—	1865.0	16.85 ± 0.08	8.9440 ± 0.0045
Exogenous CRAFT-GBM	0.1	1865.0	9.40 ± 0.09	9.1093 ± 0.0114
Exogenous CRAFT-GBM	0.3	1865.0	11.00 ± 0.10	9.0088 ± 0.0112
Exogenous CRAFT-GBM	0.5	1865.0	12.54 ± 0.03	8.9942 ± 0.0105
Exogenous CRAFT-GBM	0.7	1865.0	14.29 ± 0.05	8.9411 ± 0.0081
Exogenous CRAFT-GBM	0.9	1865.0	16.43 ± 0.06	8.9458 ± 0.0068
LightGBM-Fixed	0.1	1865.0	11.14 ± 0.03	8.9698 ± 0.0077
LightGBM-Fixed	0.3	1865.0	12.53 ± 0.04	8.9373 ± 0.0090
LightGBM-Fixed	0.5	1865.0	13.70 ± 0.08	8.9339 ± 0.0045
LightGBM-Fixed	0.7	1865.0	15.11 ± 0.08	8.9378 ± 0.0046
LightGBM-Fixed	0.9	1865.0	16.64 ± 0.08	8.9478 ± 0.0067
XGBoost	0.1	1865.0	26.14 ± 0.39	8.9251 ± 0.0029
XGBoost	0.3	1865.0	24.94 ± 0.14	8.8969 ± 0.0054
XGBoost	0.5	1865.0	25.22 ± 0.13	8.8959 ± 0.0065
XGBoost	0.7	1865.0	25.94 ± 0.14	8.9091 ± 0.0057
XGBoost	0.9	1865.0	26.39 ± 0.06	8.9051 ± 0.0072
CatBoost	0.1	1865.0	21.57 ± 0.11	8.9362 ± 0.0025
CatBoost	0.3	1865.0	30.57 ± 0.15	8.9210 ± 0.0053
CatBoost	0.5	1865.0	36.10 ± 0.40	8.9162 ± 0.0034
CatBoost	0.7	1865.0	39.04 ± 0.52	8.9165 ± 0.0048
CatBoost	0.9	1865.0	37.94 ± 0.28	8.9160 ± 0.0034
Mean ablation	0.1	1865.0	9.05 ± 0.16	9.4543 ± 0.0165
Mean ablation	0.3	1865.0	10.68 ± 0.05	9.3170 ± 0.0198
Mean ablation	0.5	1865.0	12.45 ± 0.12	9.1954 ± 0.0204
Mean ablation	0.7	1865.0	14.15 ± 0.10	9.1384 ± 0.0081
Mean ablation	0.9	1865.0	16.39 ± 0.06	9.0883 ± 0.0099
Exposure ablation	0.1	1865.0	9.35 ± 0.02	9.1031 ± 0.0071
Exposure ablation	0.3	1865.0	11.07 ± 0.06	8.9825 ± 0.0071
Exposure ablation	0.5	1865.0	12.65 ± 0.05	8.9661 ± 0.0036
Exposure ablation	0.7	1865.0	14.38 ± 0.05	8.9461 ± 0.0051
Exposure ablation	0.9	1865.0	16.42 ± 0.09	8.9444 ± 0.0094

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
APS Failure — Binary log loss (nats/observation)				
LightGBM-Full	—	571.0	3.51 ± 0.02	$0.0202 \pm 1.32 \times 10^{-4}$
LightGBM-Full (CRAFT-GBM ref.)	—	571.0	3.52 ± 0.03	$0.0202 \pm 1.32 \times 10^{-4}$
LightGBM-Full (ablation ref.)	—	571.0	3.49 ± 0.03	$0.0202 \pm 1.32 \times 10^{-4}$
Exogenous CRAFT-GBM	0.1	571.0	2.08 ± 0.02	$0.0199 \pm 1.20 \times 10^{-4}$
Exogenous CRAFT-GBM	0.3	571.0	2.45 ± 0.03	$0.0195 \pm 1.26 \times 10^{-4}$
Exogenous CRAFT-GBM	0.5	571.0	2.73 ± 0.02	$0.0196 \pm 2.40 \times 10^{-4}$
Exogenous CRAFT-GBM	0.7	571.0	3.11 ± 0.04	$0.0194 \pm 1.21 \times 10^{-4}$
Exogenous CRAFT-GBM	0.9	571.0	3.50 ± 0.02	$0.0200 \pm 1.79 \times 10^{-4}$
LightGBM-Fixed	0.1	571.0	2.31 ± 0.03	$0.0196 \pm 1.58 \times 10^{-4}$
LightGBM-Fixed	0.3	571.0	2.74 ± 0.04	$0.0191 \pm 1.59 \times 10^{-4}$
LightGBM-Fixed	0.5	571.0	3.04 ± 0.03	$0.0194 \pm 3.61 \times 10^{-4}$
LightGBM-Fixed	0.7	571.0	3.28 ± 0.03	$0.0196 \pm 9.30 \times 10^{-5}$
LightGBM-Fixed	0.9	571.0	3.45 ± 0.01	$0.0199 \pm 2.10 \times 10^{-4}$
XGBoost	0.1	571.0	4.52 ± 0.17	$0.0195 \pm 1.35 \times 10^{-4}$
XGBoost	0.3	571.0	4.74 ± 0.02	$0.0190 \pm 1.76 \times 10^{-4}$
XGBoost	0.5	571.0	5.14 ± 0.05	$0.0193 \pm 2.25 \times 10^{-4}$
XGBoost	0.7	571.0	5.47 ± 0.05	$0.0195 \pm 1.55 \times 10^{-4}$
XGBoost	0.9	571.0	5.83 ± 0.03	$0.0198 \pm 2.23 \times 10^{-4}$
CatBoost	0.1	571.0	4.50 ± 0.07	$0.0207 \pm 1.02 \times 10^{-4}$
CatBoost	0.3	571.0	6.15 ± 0.06	$0.0206 \pm 6.76 \times 10^{-5}$
CatBoost	0.5	571.0	7.18 ± 0.20	$0.0205 \pm 1.28 \times 10^{-4}$
CatBoost	0.7	571.0	7.68 ± 0.20	$0.0206 \pm 1.45 \times 10^{-4}$
CatBoost	0.9	571.0	7.06 ± 0.09	$0.0206 \pm 6.32 \times 10^{-5}$
Mean ablation	0.1	571.0	1.83 ± 0.06	$0.0260 \pm 4.32 \times 10^{-4}$
Mean ablation	0.3	571.0	2.34 ± 0.03	$0.0213 \pm 2.73 \times 10^{-4}$
Mean ablation	0.5	571.0	2.70 ± 0.02	$0.0212 \pm 1.58 \times 10^{-4}$
Mean ablation	0.7	571.0	3.15 ± 0.01	$0.0203 \pm 9.09 \times 10^{-5}$
Mean ablation	0.9	571.0	3.39 ± 0.03	$0.0201 \pm 9.82 \times 10^{-5}$
Exposure ablation	0.1	571.0	2.02 ± 0.02	$0.0206 \pm 6.29 \times 10^{-5}$
Exposure ablation	0.3	571.0	2.43 ± 0.02	$0.0193 \pm 1.33 \times 10^{-4}$
Exposure ablation	0.5	571.0	2.79 ± 0.02	$0.0192 \pm 1.40 \times 10^{-4}$
Exposure ablation	0.7	571.0	3.11 ± 0.03	$0.0193 \pm 1.26 \times 10^{-4}$
Exposure ablation	0.9	571.0	3.36 ± 0.03	$0.0198 \pm 1.60 \times 10^{-4}$

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
KDD 2009 — Binary log loss (nats/observation)				
LightGBM-Full	—	93.0	1.22 ± 0.01	$0.0770 \pm 1.82 \times 10^{-4}$
LightGBM-Full (CRAFT-GBM ref.)	—	93.0	1.20 ± 0.00	$0.0770 \pm 1.82 \times 10^{-4}$
LightGBM-Full (ablation ref.)	—	93.0	1.21 ± 0.00	$0.0770 \pm 1.82 \times 10^{-4}$
Exogenous CRAFT-GBM	0.1	93.0	0.99 ± 0.00	$0.0804 \pm 5.60 \times 10^{-5}$
Exogenous CRAFT-GBM	0.3	93.0	1.11 ± 0.03	$0.0755 \pm 1.91 \times 10^{-4}$
Exogenous CRAFT-GBM	0.5	93.0	1.17 ± 0.01	$0.0759 \pm 1.04 \times 10^{-4}$
Exogenous CRAFT-GBM	0.7	93.0	1.24 ± 0.01	$0.0762 \pm 1.20 \times 10^{-4}$
Exogenous CRAFT-GBM	0.9	93.0	1.26 ± 0.00	$0.0769 \pm 4.96 \times 10^{-5}$
LightGBM-Fixed	0.1	93.0	1.02 ± 0.10	$0.0809 \pm 2.48 \times 10^{-4}$
LightGBM-Fixed	0.3	93.0	1.08 ± 0.05	$0.0775 \pm 4.74 \times 10^{-4}$
LightGBM-Fixed	0.5	93.0	1.13 ± 0.01	$0.0767 \pm 2.56 \times 10^{-4}$
LightGBM-Fixed	0.7	93.0	1.17 ± 0.01	$0.0765 \pm 1.19 \times 10^{-4}$
LightGBM-Fixed	0.9	93.0	1.20 ± 0.01	$0.0770 \pm 2.97 \times 10^{-4}$
XGBoost	0.1	93.0	2.23 ± 0.04	$0.0835 \pm 7.47 \times 10^{-4}$
XGBoost	0.3	93.0	3.30 ± 0.05	$0.0895 \pm 7.61 \times 10^{-4}$
XGBoost	0.5	93.0	4.22 ± 0.04	$0.0914 \pm 5.02 \times 10^{-4}$
XGBoost	0.7	93.0	4.82 ± 0.07	$0.0916 \pm 4.75 \times 10^{-4}$
XGBoost	0.9	93.0	5.04 ± 0.04	$0.0915 \pm 5.27 \times 10^{-4}$
CatBoost	0.1	93.0	2.59 ± 0.12	$0.0879 \pm 2.97 \times 10^{-4}$
CatBoost	0.3	93.0	2.80 ± 0.19	$0.0861 \pm 3.95 \times 10^{-4}$
CatBoost	0.5	93.0	2.95 ± 0.06	$0.0851 \pm 4.95 \times 10^{-4}$
CatBoost	0.7	93.0	3.24 ± 0.32	$0.0846 \pm 7.21 \times 10^{-4}$
CatBoost	0.9	93.0	3.28 ± 0.19	$0.0845 \pm 1.86 \times 10^{-4}$
Mean ablation	0.1	93.0	0.99 ± 0.01	$0.0765 \pm 9.48 \times 10^{-5}$
Mean ablation	0.3	93.0	1.08 ± 0.01	$0.0767 \pm 8.35 \times 10^{-5}$
Mean ablation	0.5	93.0	1.13 ± 0.01	$0.0772 \pm 1.72 \times 10^{-4}$
Mean ablation	0.7	93.0	1.19 ± 0.00	$0.0770 \pm 9.75 \times 10^{-5}$
Mean ablation	0.9	93.0	1.20 ± 0.00	$0.0770 \pm 7.91 \times 10^{-5}$
Exposure ablation	0.1	93.0	0.93 ± 0.00	$0.0804 \pm 5.60 \times 10^{-5}$
Exposure ablation	0.3	93.0	1.02 ± 0.00	$0.0768 \pm 1.41 \times 10^{-4}$
Exposure ablation	0.5	93.0	1.09 ± 0.01	$0.0763 \pm 9.70 \times 10^{-5}$
Exposure ablation	0.7	93.0	1.15 ± 0.00	$0.0758 \pm 8.51 \times 10^{-5}$
Exposure ablation	0.9	93.0	1.19 ± 0.00	$0.0767 \pm 1.07 \times 10^{-4}$

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
Gisette — Binary log loss (nats/observation)				
LightGBM-Full	—	1652.0	62.92 ± 0.34	$0.0697 \pm 4.52\times10^{-4}$
LightGBM-Full (CRAFT-GBM ref.)	—	1652.0	62.85 ± 0.30	$0.0697 \pm 4.52\times10^{-4}$
LightGBM-Full (ablation ref.)	—	1652.0	62.92 ± 0.16	$0.0697 \pm 4.52\times10^{-4}$
Exogenous CRAFT-GBM	0.1	1652.0	13.79 ± 0.04	$0.0614 \pm 4.84\times10^{-4}$
Exogenous CRAFT-GBM	0.3	1652.0	30.17 ± 0.08	$0.0638 \pm 9.50\times10^{-4}$
Exogenous CRAFT-GBM	0.5	1652.0	52.82 ± 0.14	$0.0666 \pm 8.28\times10^{-4}$
Exogenous CRAFT-GBM	0.7	1652.0	61.80 ± 0.09	$0.0676 \pm 4.17\times10^{-4}$
Exogenous CRAFT-GBM	0.9	1652.0	64.08 ± 0.13	$0.0691 \pm 6.02\times10^{-4}$
LightGBM-Fixed	0.1	1652.0	18.17 ± 0.03	0.0602 ± 0.0012
LightGBM-Fixed	0.3	1652.0	30.73 ± 0.06	$0.0625 \pm 3.36\times10^{-4}$
LightGBM-Fixed	0.5	1652.0	41.10 ± 0.06	$0.0644 \pm 7.86\times10^{-4}$
LightGBM-Fixed	0.7	1652.0	60.23 ± 0.16	$0.0665 \pm 3.21\times10^{-4}$
LightGBM-Fixed	0.9	1652.0	62.21 ± 0.14	$0.0680 \pm 4.95\times10^{-4}$
XGBoost	0.1	1652.0	98.91 ± 0.18	$0.0615 \pm 3.75\times10^{-4}$
XGBoost	0.3	1652.0	107.15 ± 0.19	$0.0626 \pm 6.76\times10^{-4}$
XGBoost	0.5	1652.0	115.54 ± 0.77	$0.0650 \pm 7.63\times10^{-4}$
XGBoost	0.7	1652.0	123.74 ± 0.50	$0.0673 \pm 3.93\times10^{-4}$
XGBoost	0.9	1652.0	132.24 ± 0.50	0.0717 ± 0.0012
CatBoost	0.1	1652.0	61.23 ± 0.17	$0.0725 \pm 7.84\times10^{-4}$
CatBoost	0.3	1652.0	148.77 ± 3.66	0.0725 ± 0.0012
CatBoost	0.5	1652.0	240.63 ± 0.37	$0.0722 \pm 8.08\times10^{-4}$
CatBoost	0.7	1652.0	268.39 ± 0.48	0.0722 ± 0.0011
CatBoost	0.9	1652.0	262.38 ± 0.55	$0.0714 \pm 6.54\times10^{-4}$
Mean ablation	0.1	1652.0	14.98 ± 0.08	0.0612 ± 0.0012
Mean ablation	0.3	1652.0	29.80 ± 0.08	$0.0653 \pm 6.72\times10^{-4}$
Mean ablation	0.5	1652.0	40.99 ± 0.08	0.0677 ± 0.0010
Mean ablation	0.7	1652.0	60.52 ± 0.17	$0.0686 \pm 6.85\times10^{-4}$
Mean ablation	0.9	1652.0	63.19 ± 0.18	$0.0696 \pm 9.25\times10^{-4}$
Exposure ablation	0.1	1652.0	10.17 ± 0.07	$0.0642 \pm 6.69\times10^{-4}$
Exposure ablation	0.3	1652.0	22.02 ± 0.09	$0.0632 \pm 7.20\times10^{-4}$
Exposure ablation	0.5	1652.0	33.48 ± 0.10	$0.0651 \pm 7.25\times10^{-4}$
Exposure ablation	0.7	1652.0	58.91 ± 0.13	$0.0661 \pm 7.22\times10^{-4}$
Exposure ablation	0.9	1652.0	61.79 ± 0.06	$0.0683 \pm 8.51\times10^{-4}$

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
Jannis — Multiclass log loss (nats/observation)				
LightGBM-Full	—	862.0	7.27 ± 0.12	$0.6963 \pm 6.16 \times 10^{-4}$
LightGBM-Full (CRAFT-GBM ref.)	—	862.0	7.22 ± 0.05	$0.6963 \pm 6.16 \times 10^{-4}$
LightGBM-Full (ablation ref.)	—	862.0	7.29 ± 0.10	$0.6963 \pm 6.16 \times 10^{-4}$
Exogenous CRAFT-GBM	0.1	862.0	4.96 ± 0.03	0.7521 ± 0.0015
Exogenous CRAFT-GBM	0.3	862.0	5.60 ± 0.08	$0.6822 \pm 9.10 \times 10^{-4}$
Exogenous CRAFT-GBM	0.5	862.0	6.06 ± 0.12	$0.6767 \pm 3.91 \times 10^{-4}$
Exogenous CRAFT-GBM	0.7	862.0	6.65 ± 0.13	$0.6880 \pm 2.95 \times 10^{-4}$
Exogenous CRAFT-GBM	0.9	862.0	7.06 ± 0.04	$0.6946 \pm 4.26 \times 10^{-4}$
LightGBM-Fixed	0.1	862.0	5.11 ± 0.15	0.7550 ± 0.0026
LightGBM-Fixed	0.3	862.0	5.86 ± 0.09	$0.7042 \pm 4.51 \times 10^{-4}$
LightGBM-Fixed	0.5	862.0	6.37 ± 0.12	$0.6986 \pm 8.25 \times 10^{-4}$
LightGBM-Fixed	0.7	862.0	6.76 ± 0.14	$0.6971 \pm 3.43 \times 10^{-4}$
LightGBM-Fixed	0.9	862.0	7.15 ± 0.15	$0.6961 \pm 2.91 \times 10^{-4}$
XGBoost	0.1	862.0	9.69 ± 0.04	0.7601 ± 0.0014
XGBoost	0.3	862.0	10.25 ± 0.03	$0.6994 \pm 1.38 \times 10^{-4}$
XGBoost	0.5	862.0	10.66 ± 0.03	$0.6932 \pm 2.43 \times 10^{-4}$
XGBoost	0.7	862.0	11.15 ± 0.09	$0.6928 \pm 1.76 \times 10^{-4}$
XGBoost	0.9	862.0	11.54 ± 0.04	$0.6937 \pm 1.65 \times 10^{-4}$
CatBoost	0.1	862.0	7.43 ± 0.10	$0.7344 \pm 5.21 \times 10^{-4}$
CatBoost	0.3	862.0	9.19 ± 0.05	$0.7217 \pm 3.72 \times 10^{-4}$
CatBoost	0.5	862.0	10.86 ± 0.14	$0.7192 \pm 3.23 \times 10^{-4}$
CatBoost	0.7	862.0	11.61 ± 0.31	$0.7183 \pm 3.86 \times 10^{-4}$
CatBoost	0.9	862.0	11.81 ± 0.24	$0.7178 \pm 3.59 \times 10^{-4}$
Mean ablation	0.1	862.0	4.84 ± 0.07	0.7504 ± 0.0033
Mean ablation	0.3	862.0	5.26 ± 0.13	$0.7093 \pm 4.15 \times 10^{-4}$
Mean ablation	0.5	862.0	5.85 ± 0.12	$0.6757 \pm 7.98 \times 10^{-4}$
Mean ablation	0.7	862.0	6.57 ± 0.09	$0.6876 \pm 2.88 \times 10^{-4}$
Mean ablation	0.9	862.0	6.88 ± 0.10	$0.6945 \pm 3.59 \times 10^{-4}$
Exposure ablation	0.1	862.0	4.80 ± 0.09	$0.7324 \pm 4.35 \times 10^{-4}$
Exposure ablation	0.3	862.0	5.40 ± 0.18	$0.7039 \pm 9.19 \times 10^{-4}$
Exposure ablation	0.5	862.0	6.08 ± 0.06	$0.6976 \pm 4.39 \times 10^{-4}$
Exposure ablation	0.7	862.0	6.47 ± 0.04	$0.6957 \pm 4.77 \times 10^{-4}$
Exposure ablation	0.9	862.0	6.90 ± 0.21	$0.6962 \pm 3.90 \times 10^{-4}$

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
HAR — Multiclass log loss (nats/observation)				
LightGBM-Full	—	449.0	4.66 ± 0.06	0.1565 ± 0.0030
LightGBM-Full (CRAFT-GBM ref.)	—	449.0	4.64 ± 0.05	0.1565 ± 0.0030
LightGBM-Full (ablation ref.)	—	449.0	4.57 ± 0.02	0.1565 ± 0.0030
Exogenous CRAFT-GBM	0.1	449.0	2.41 ± 0.04	0.1475 ± 0.0061
Exogenous CRAFT-GBM	0.3	449.0	2.90 ± 0.04	0.1610 ± 0.0033
Exogenous CRAFT-GBM	0.5	449.0	3.40 ± 0.04	0.1734 ± 0.0061
Exogenous CRAFT-GBM	0.7	449.0	3.98 ± 0.04	0.1674 ± 0.0033
Exogenous CRAFT-GBM	0.9	449.0	4.62 ± 0.05	0.1550 ± 0.0026
LightGBM-Fixed	0.1	449.0	2.61 ± 0.04	0.1355 ± 0.0117
LightGBM-Fixed	0.3	449.0	3.13 ± 0.03	0.1379 ± 0.0034
LightGBM-Fixed	0.5	449.0	3.67 ± 0.04	0.1482 ± 0.0046
LightGBM-Fixed	0.7	449.0	4.09 ± 0.03	0.1532 ± 0.0019
LightGBM-Fixed	0.9	449.0	4.49 ± 0.04	0.1524 ± 0.0024
XGBoost	0.1	449.0	13.90 ± 0.07	0.1316 ± 0.0055
XGBoost	0.3	449.0	13.40 ± 0.03	0.1485 ± 0.0057
XGBoost	0.5	449.0	13.65 ± 0.04	0.1610 ± 0.0060
XGBoost	0.7	449.0	14.04 ± 0.02	0.1654 ± 0.0027
XGBoost	0.9	449.0	14.55 ± 0.14	0.1666 ± 0.0024
CatBoost	0.1	449.0	17.78 ± 0.13	0.1577 ± 0.0036
CatBoost	0.3	449.0	41.31 ± 0.11	0.1674 ± 0.0053
CatBoost	0.5	449.0	58.95 ± 0.35	0.1709 ± 0.0057
CatBoost	0.7	449.0	65.69 ± 0.44	0.1716 ± 0.0078
CatBoost	0.9	449.0	65.15 ± 0.17	0.1735 ± 0.0063
Mean ablation	0.1	449.0	2.16 ± 0.04	0.5044 ± 0.0105
Mean ablation	0.3	449.0	3.10 ± 0.02	0.3751 ± 0.0045
Mean ablation	0.5	449.0	3.83 ± 0.03	0.2693 ± 0.0043
Mean ablation	0.7	449.0	3.81 ± 0.04	0.2583 ± 0.0033
Mean ablation	0.9	449.0	4.41 ± 0.02	0.1640 ± 0.0024
Exposure ablation	0.1	449.0	2.35 ± 0.04	0.1362 ± 0.0058
Exposure ablation	0.3	449.0	3.06 ± 0.04	0.1412 ± 0.0022
Exposure ablation	0.5	449.0	3.56 ± 0.03	0.1521 ± 0.0039
Exposure ablation	0.7	449.0	4.06 ± 0.03	0.1511 ± 0.0045
Exposure ablation	0.9	449.0	4.51 ± 0.03	0.1566 ± 0.0020

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
CT Scan — RMSE (target units)				
LightGBM-Full	—	1133.6	11.65 [6.77, 17.22]	6.0271 [4.7394, 6.6746]
LightGBM-Full (CRAFT-GBM ref.)	—	1133.6	11.66 [6.80, 17.31]	6.0271 [4.7394, 6.6746]
LightGBM-Full (ablation ref.)	—	1133.6	11.61 [6.77, 17.19]	6.0271 [4.7394, 6.6746]
Exogenous CRAFT-GBM	0.1	1133.6	3.91 [2.58, 5.50]	6.1973 [4.6741, 6.9082]
Exogenous CRAFT-GBM	0.3	1133.6	5.95 [3.86, 8.67]	5.8415 [4.4166, 6.6305]
Exogenous CRAFT-GBM	0.5	1133.6	7.85 [4.97, 11.33]	5.8922 [4.3699, 6.6348]
Exogenous CRAFT-GBM	0.7	1133.6	10.93 [6.49, 16.14]	5.9481 [4.4811, 6.6416]
Exogenous CRAFT-GBM	0.9	1133.6	11.61 [6.78, 17.08]	6.0239 [4.7292, 6.7180]
LightGBM-Fixed	0.1	1133.6	5.22 [3.31, 7.47]	5.9570 [4.2750, 6.9592]
LightGBM-Fixed	0.3	1133.6	6.88 [4.30, 9.98]	5.8224 [4.1977, 6.7020]
LightGBM-Fixed	0.5	1133.6	8.37 [5.16, 12.23]	5.8768 [4.4525, 6.7008]
LightGBM-Fixed	0.7	1133.6	11.26 [6.63, 16.62]	5.9766 [4.4187, 6.6770]
LightGBM-Fixed	0.9	1133.6	11.50 [6.69, 16.91]	5.9779 [4.6965, 6.5546]
XGBoost	0.1	1133.6	11.48 [7.02, 16.59]	5.9310 [4.3268, 6.7232]
XGBoost	0.3	1133.6	12.70 [8.07, 18.43]	5.8669 [4.2006, 6.6701]
XGBoost	0.5	1133.6	13.21 [7.99, 19.35]	5.9068 [4.4750, 6.7424]
XGBoost	0.7	1133.6	14.42 [8.77, 21.33]	5.9725 [4.6502, 6.7465]
XGBoost	0.9	1133.6	15.58 [9.48, 22.92]	6.0420 [4.8552, 6.8699]
CatBoost	0.1	1133.6	7.26 [4.86, 10.04]	6.2353 [4.7208, 7.0001]
CatBoost	0.3	1133.6	12.24 [7.67, 18.87]	6.1669 [4.9025, 6.8974]
CatBoost	0.5	1133.6	16.50 [11.23, 23.02]	6.1774 [4.7892, 6.9687]
CatBoost	0.7	1133.6	18.64 [11.28, 28.93]	6.1575 [4.9248, 6.9048]
CatBoost	0.9	1133.6	17.29 [10.73, 25.21]	6.1831 [4.9640, 7.0183]
Mean ablation	0.1	1133.6	3.65 [2.54, 4.75]	7.3850 [6.2800, 8.7861]
Mean ablation	0.3	1133.6	5.57 [3.68, 7.91]	6.5659 [5.0137, 7.6150]
Mean ablation	0.5	1133.6	7.69 [4.66, 10.97]	6.2275 [4.8044, 7.0996]
Mean ablation	0.7	1133.6	10.75 [6.32, 15.90]	5.9488 [4.5501, 6.6749]
Mean ablation	0.9	1133.6	11.59 [6.59, 16.59]	6.0854 [4.7297, 6.7443]
Exposure ablation	0.1	1133.6	3.67 [2.45, 5.04]	6.4033 [4.7410, 7.2426]
Exposure ablation	0.3	1133.6	5.40 [3.54, 7.70]	6.0037 [4.4365, 6.7595]
Exposure ablation	0.5	1133.6	7.29 [4.54, 10.59]	6.0590 [4.2883, 7.1772]
Exposure ablation	0.7	1133.6	10.59 [6.14, 15.68]	6.0294 [4.5875, 6.7078]
Exposure ablation	0.9	1133.6	11.27 [6.58, 16.66]	5.9799 [4.6135, 6.7541]

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
Sparse 1 — Binary log loss (nats/observation)				
LightGBM-Full	—	1537.8	69.90 ± 19.69	0.4976 ± 0.0017
LightGBM-Full (CRAFT-GBM ref.)	—	1537.8	69.87 ± 19.62	0.4976 ± 0.0017
LightGBM-Full (ablation ref.)	—	1537.8	69.79 ± 19.67	0.4976 ± 0.0017
Exogenous CRAFT-GBM	0.1	1537.8	12.84 ± 3.46	0.4898 ± 0.0014
Exogenous CRAFT-GBM	0.3	1537.8	24.96 ± 6.84	$0.4939 \pm 9.85 \times 10^{-4}$
Exogenous CRAFT-GBM	0.5	1537.8	37.46 ± 10.49	0.4956 ± 0.0013
Exogenous CRAFT-GBM	0.7	1537.8	63.83 ± 17.98	0.4964 ± 0.0013
Exogenous CRAFT-GBM	0.9	1537.8	68.43 ± 19.20	0.4974 ± 0.0014
LightGBM-Fixed	0.1	1537.8	31.31 ± 8.66	0.5067 ± 0.0016
LightGBM-Fixed	0.3	1537.8	39.26 ± 11.00	0.4990 ± 0.0012
LightGBM-Fixed	0.5	1537.8	47.30 ± 13.27	0.4985 ± 0.0013
LightGBM-Fixed	0.7	1537.8	69.00 ± 19.45	0.4980 ± 0.0013
LightGBM-Fixed	0.9	1537.8	69.60 ± 19.54	0.4975 ± 0.0015
XGBoost	0.1	1537.8	121.58 ± 33.21	0.5092 ± 0.0036
XGBoost	0.3	1537.8	128.01 ± 35.14	0.5103 ± 0.0035
XGBoost	0.5	1537.8	134.93 ± 36.87	0.5112 ± 0.0037
XGBoost	0.7	1537.8	142.31 ± 39.23	0.5119 ± 0.0038
XGBoost	0.9	1537.8	149.93 ± 41.24	0.5122 ± 0.0037
CatBoost	0.1	1537.8	65.76 ± 18.24	0.5078 ± 0.0039
CatBoost	0.3	1537.8	144.17 ± 40.91	0.5080 ± 0.0038
CatBoost	0.5	1537.8	197.36 ± 56.39	0.5082 ± 0.0037
CatBoost	0.7	1537.8	218.74 ± 62.13	0.5083 ± 0.0038
CatBoost	0.9	1537.8	218.80 ± 61.85	0.5085 ± 0.0037
Mean ablation	0.1	1537.8	13.28 ± 3.48	0.4890 ± 0.0011
Mean ablation	0.3	1537.8	24.95 ± 6.82	0.4939 ± 0.0014
Mean ablation	0.5	1537.8	37.02 ± 10.33	0.4953 ± 0.0015
Mean ablation	0.7	1537.8	63.49 ± 17.92	0.4967 ± 0.0017
Mean ablation	0.9	1537.8	68.01 ± 19.19	0.4972 ± 0.0015
Exposure ablation	0.1	1537.8	10.53 ± 2.65	0.5057 ± 0.0018
Exposure ablation	0.3	1537.8	22.24 ± 6.04	0.4995 ± 0.0013
Exposure ablation	0.5	1537.8	34.54 ± 9.63	0.4986 ± 0.0011
Exposure ablation	0.7	1537.8	62.70 ± 17.67	0.4977 ± 0.0014
Exposure ablation	0.9	1537.8	67.26 ± 18.97	0.4975 ± 0.0012

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
Sparse 2 — Binary log loss (nats/observation)				
LightGBM-Full	—	1339.4	4.23 ± 1.56	0.4656 ± 0.0016
LightGBM-Full (CRAFT-GBM ref.)	—	1339.4	4.23 ± 1.55	0.4656 ± 0.0016
LightGBM-Full (ablation ref.)	—	1339.4	4.23 ± 1.55	0.4656 ± 0.0016
Exogenous CRAFT-GBM	0.1	1339.4	3.86 ± 1.30	0.4636 ± 0.0016
Exogenous CRAFT-GBM	0.3	1339.4	4.37 ± 1.56	0.4646 ± 0.0015
Exogenous CRAFT-GBM	0.5	1339.4	4.47 ± 1.59	0.4652 ± 0.0015
Exogenous CRAFT-GBM	0.7	1339.4	4.55 ± 1.61	0.4654 ± 0.0018
Exogenous CRAFT-GBM	0.9	1339.4	4.56 ± 1.67	0.4654 ± 0.0017
LightGBM-Fixed	0.1	1339.4	3.94 ± 1.39	0.4754 ± 0.0026
LightGBM-Fixed	0.3	1339.4	4.13 ± 1.47	0.4675 ± 0.0022
LightGBM-Fixed	0.5	1339.4	4.18 ± 1.53	0.4665 ± 0.0019
LightGBM-Fixed	0.7	1339.4	4.24 ± 1.51	0.4660 ± 0.0017
LightGBM-Fixed	0.9	1339.4	4.25 ± 1.53	0.4656 ± 0.0016
XGBoost	0.1	1339.4	33.32 ± 12.30	0.4705 ± 0.0029
XGBoost	0.3	1339.4	33.85 ± 12.38	0.4690 ± 0.0020
XGBoost	0.5	1339.4	33.52 ± 12.31	0.4676 ± 0.0015
XGBoost	0.7	1339.4	33.75 ± 12.29	0.4670 ± 0.0014
XGBoost	0.9	1339.4	34.14 ± 12.02	0.4672 ± 0.0013
CatBoost	0.1	1339.4	27.22 ± 11.14	0.4668 ± 0.0021
CatBoost	0.3	1339.4	28.43 ± 11.69	0.4660 ± 0.0017
CatBoost	0.5	1339.4	27.67 ± 11.39	0.4658 ± 0.0015
CatBoost	0.7	1339.4	28.25 ± 11.33	0.4659 ± 0.0014
CatBoost	0.9	1339.4	27.76 ± 11.40	0.4659 ± 0.0014
Mean ablation	0.1	1339.4	3.67 ± 1.25	$0.4633 \pm 9.94 \times 10^{-4}$
Mean ablation	0.3	1339.4	4.20 ± 1.50	0.4646 ± 0.0014
Mean ablation	0.5	1339.4	4.22 ± 1.55	0.4651 ± 0.0016
Mean ablation	0.7	1339.4	4.31 ± 1.56	0.4655 ± 0.0018
Mean ablation	0.9	1339.4	4.38 ± 1.53	0.4655 ± 0.0018
Exposure ablation	0.1	1339.4	3.86 ± 1.37	0.4766 ± 0.0037
Exposure ablation	0.3	1339.4	4.16 ± 1.50	0.4687 ± 0.0023
Exposure ablation	0.5	1339.4	4.20 ± 1.53	0.4675 ± 0.0019
Exposure ablation	0.7	1339.4	4.29 ± 1.55	0.4661 ± 0.0017
Exposure ablation	0.9	1339.4	4.37 ± 1.60	0.4656 ± 0.0018

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
Dense 1 — RMSE (target units)				
LightGBM-Full	—	1659.2	76.15 ± 6.14	1.0449 ± 0.0026
LightGBM-Full (CRAFT-GBM ref.)	—	1659.2	75.42 ± 6.08	1.0449 ± 0.0026
LightGBM-Full (ablation ref.)	—	1659.2	75.64 ± 6.00	1.0449 ± 0.0026
Exogenous CRAFT-GBM	0.1	1659.2	11.04 ± 0.76	1.0407 ± 0.0031
Exogenous CRAFT-GBM	0.3	1659.2	24.81 ± 1.80	1.0370 ± 0.0023
Exogenous CRAFT-GBM	0.5	1659.2	40.10 ± 3.02	1.0391 ± 0.0025
Exogenous CRAFT-GBM	0.7	1659.2	55.40 ± 4.33	1.0423 ± 0.0024
Exogenous CRAFT-GBM	0.9	1659.2	70.21 ± 5.53	1.0435 ± 0.0015
LightGBM-Fixed	0.1	1659.2	17.02 ± 1.54	2.0715 ± 0.2535
LightGBM-Fixed	0.3	1659.2	30.25 ± 2.34	1.1259 ± 0.0134
LightGBM-Fixed	0.5	1659.2	44.62 ± 3.30	1.0518 ± 0.0039
LightGBM-Fixed	0.7	1659.2	58.40 ± 4.52	1.0392 ± 0.0028
LightGBM-Fixed	0.9	1659.2	70.13 ± 5.67	1.0410 ± 0.0018
XGBoost	0.1	1659.2	222.05 ± 20.43	1.7724 ± 0.1324
XGBoost	0.3	1659.2	247.66 ± 19.81	1.1253 ± 0.0235
XGBoost	0.5	1659.2	253.64 ± 20.38	1.0454 ± 0.0059
XGBoost	0.7	1659.2	259.86 ± 21.50	1.0373 ± 0.0026
XGBoost	0.9	1659.2	267.46 ± 22.91	1.0393 ± 0.0021
CatBoost	0.1	1659.2	57.46 ± 4.75	1.3006 ± 0.0598
CatBoost	0.3	1659.2	133.74 ± 12.83	1.0303 ± 0.0038
CatBoost	0.5	1659.2	191.34 ± 17.55	1.0155 ± 0.0015
CatBoost	0.7	1659.2	219.75 ± 20.47	1.0126 ± 0.0012
CatBoost	0.9	1659.2	224.03 ± 21.77	1.0115 ± 0.0012
Mean ablation	0.1	1659.2	10.74 ± 0.71	1.0383 ± 0.0016
Mean ablation	0.3	1659.2	24.04 ± 1.76	1.0416 ± 0.0019
Mean ablation	0.5	1659.2	38.96 ± 2.97	1.0428 ± 0.0018
Mean ablation	0.7	1659.2	53.86 ± 4.18	1.0441 ± 0.0012
Mean ablation	0.9	1659.2	68.56 ± 5.55	1.0443 ± 0.0023
Exposure ablation	0.1	1659.2	11.08 ± 0.80	1.8845 ± 0.0694
Exposure ablation	0.3	1659.2	24.47 ± 1.68	1.6814 ± 0.0023
Exposure ablation	0.5	1659.2	39.33 ± 2.97	1.5483 ± 0.2916
Exposure ablation	0.7	1659.2	54.07 ± 4.10	1.0556 ± 0.0093
Exposure ablation	0.9	1659.2	67.61 ± 5.46	1.0429 ± 0.0024

Continued on next page

Table 10 (continued)

Model	f	Mean rounds	Training (s)	Test loss
Dense 2 — Binary log loss (nats/observation)				
LightGBM-Full	—	1812.8	77.92 ± 2.28	0.3774 ± 0.0178
LightGBM-Full (CRAFT-GBM ref.)	—	1812.8	77.73 ± 2.21	0.3774 ± 0.0178
LightGBM-Full (ablation ref.)	—	1812.8	77.65 ± 2.45	0.3774 ± 0.0178
Exogenous CRAFT-GBM	0.1	1812.8	12.08 ± 0.47	0.3694 ± 0.0177
Exogenous CRAFT-GBM	0.3	1812.8	26.27 ± 0.87	0.3735 ± 0.0179
Exogenous CRAFT-GBM	0.5	1812.8	41.77 ± 1.39	0.3752 ± 0.0184
Exogenous CRAFT-GBM	0.7	1812.8	57.48 ± 1.72	0.3763 ± 0.0179
Exogenous CRAFT-GBM	0.9	1812.8	72.31 ± 1.98	0.3770 ± 0.0179
LightGBM-Fixed	0.1	1812.8	19.32 ± 0.91	0.3981 ± 0.0166
LightGBM-Fixed	0.3	1812.8	31.52 ± 1.04	0.3813 ± 0.0173
LightGBM-Fixed	0.5	1812.8	45.64 ± 1.40	0.3783 ± 0.0174
LightGBM-Fixed	0.7	1812.8	59.42 ± 1.82	0.3779 ± 0.0179
LightGBM-Fixed	0.9	1812.8	71.79 ± 2.32	0.3776 ± 0.0177
XGBoost	0.1	1812.8	235.93 ± 10.44	0.3800 ± 0.0159
XGBoost	0.3	1812.8	245.23 ± 9.30	0.3671 ± 0.0169
XGBoost	0.5	1812.8	251.84 ± 10.14	0.3647 ± 0.0178
XGBoost	0.7	1812.8	263.78 ± 9.31	0.3638 ± 0.0182
XGBoost	0.9	1812.8	275.24 ± 10.40	0.3637 ± 0.0178
CatBoost	0.1	1812.8	67.01 ± 2.67	0.3844 ± 0.0158
CatBoost	0.3	1812.8	149.81 ± 5.08	0.3757 ± 0.0173
CatBoost	0.5	1812.8	213.03 ± 8.39	0.3741 ± 0.0174
CatBoost	0.7	1812.8	246.24 ± 11.86	0.3733 ± 0.0175
CatBoost	0.9	1812.8	252.67 ± 11.37	0.3725 ± 0.0177
Mean ablation	0.1	1812.8	11.52 ± 0.46	0.3671 ± 0.0180
Mean ablation	0.3	1812.8	25.54 ± 0.87	0.3738 ± 0.0180
Mean ablation	0.5	1812.8	40.63 ± 1.36	0.3754 ± 0.0183
Mean ablation	0.7	1812.8	56.26 ± 2.01	0.3768 ± 0.0183
Mean ablation	0.9	1812.8	70.80 ± 2.39	0.3773 ± 0.0177
Exposure ablation	0.1	1812.8	12.63 ± 0.46	0.4054 ± 0.0175
Exposure ablation	0.3	1812.8	25.51 ± 0.97	0.3827 ± 0.0172
Exposure ablation	0.5	1812.8	39.83 ± 1.16	0.3813 ± 0.0174
Exposure ablation	0.7	1812.8	54.70 ± 1.69	0.3778 ± 0.0179
Exposure ablation	0.9	1812.8	68.93 ± 2.19	0.3775 ± 0.0181

Table 11: RQ2: paired percentage changes relative to LightGBM-Full. Entries apply $100(r-1)$ to geometric-mean ratios and pointwise paired 95% interval endpoints, whereas CT brackets are descriptive fold ranges. Negative values are favorable. Exogenous CRAFT-GBM and the ablations each use their same-campaign Full reference, while the remaining models use the original Full reference.

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
HIGGS					
Exogenous CRAFT-GBM	0.1	-22.00	[-22.72, -21.27]	9.66	[8.94, 10.38]
Exogenous CRAFT-GBM	0.3	-17.42	[-17.73, -17.11]	4.07	[3.53, 4.62]
Exogenous CRAFT-GBM	0.5	-13.77	[-14.36, -13.18]	2.60	[2.48, 2.72]
Exogenous CRAFT-GBM	0.7	-5.57	[-5.98, -5.17]	0.63	[0.49, 0.76]
Exogenous CRAFT-GBM	0.9	-2.83	[-3.09, -2.57]	0.49	[0.34, 0.65]
LightGBM-Fixed	0.1	-14.81	[-15.38, -14.23]	5.82	[5.72, 5.92]
LightGBM-Fixed	0.3	-11.68	[-12.54, -10.82]	1.89	[1.69, 2.10]
LightGBM-Fixed	0.5	-7.06	[-7.32, -6.79]	0.72	[0.67, 0.77]
LightGBM-Fixed	0.7	-2.61	[-3.04, -2.18]	0.31	[0.27, 0.35]
LightGBM-Fixed	0.9	-1.17	[-1.48, -0.86]	0.13	[0.06, 0.19]
XGBoost	0.1	35.05	[34.57, 35.53]	8.28	[8.07, 8.50]
XGBoost	0.3	38.03	[37.55, 38.51]	0.67	[0.61, 0.74]
XGBoost	0.5	36.45	[35.97, 36.94]	-0.58	[-0.64, -0.52]
XGBoost	0.7	36.41	[34.66, 38.18]	-0.96	[-1.00, -0.93]
XGBoost	0.9	36.14	[34.53, 37.78]	-1.24	[-1.32, -1.16]
CatBoost	0.1	33.29	[32.46, 34.13]	3.39	[3.33, 3.44]
CatBoost	0.3	65.46	[65.06, 65.87]	1.30	[1.28, 1.32]
CatBoost	0.5	85.92	[85.06, 86.79]	0.93	[0.88, 0.97]
CatBoost	0.7	95.87	[94.25, 97.51]	0.77	[0.73, 0.81]
CatBoost	0.9	96.96	[95.85, 98.09]	0.70	[0.66, 0.73]
Mean ablation	0.1	-22.21	[-22.46, -21.96]	9.49	[8.75, 10.24]
Mean ablation	0.3	-18.59	[-19.20, -17.98]	7.81	[7.33, 8.30]
Mean ablation	0.5	-9.50	[-10.38, -8.61]	3.00	[2.95, 3.05]
Mean ablation	0.7	-3.47	[-3.60, -3.34]	1.55	[1.50, 1.59]
Mean ablation	0.9	-2.73	[-3.17, -2.29]	0.30	[0.26, 0.34]
Exposure ablation	0.1	-19.63	[-20.00, -19.25]	8.91	[8.87, 8.95]
Exposure ablation	0.3	-15.61	[-15.89, -15.33]	3.81	[3.76, 3.85]
Exposure ablation	0.5	-10.82	[-10.96, -10.68]	3.13	[3.09, 3.16]
Exposure ablation	0.7	-4.05	[-4.31, -3.79]	0.15	[0.06, 0.25]
Exposure ablation	0.9	-1.93	[-2.54, -1.33]	0.09	[0.03, 0.15]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
Epsilon					
Exogenous CRAFT-GBM	0.1	-84.15	[-84.20, -84.10]	-2.33	[-2.60, -2.06]
Exogenous CRAFT-GBM	0.3	-64.66	[-64.70, -64.62]	-1.40	[-1.81, -0.99]
Exogenous CRAFT-GBM	0.5	-46.07	[-46.30, -45.84]	-1.00	[-1.28, -0.71]
Exogenous CRAFT-GBM	0.7	-27.38	[-27.55, -27.22]	-0.63	[-0.91, -0.36]
Exogenous CRAFT-GBM	0.9	-8.26	[-8.50, -8.02]	-0.35	[-0.67, -0.02]
LightGBM-Fixed	0.1	-72.97	[-73.02, -72.93]	0.59	[0.25, 0.93]
LightGBM-Fixed	0.3	-55.84	[-55.97, -55.72]	-0.23	[-0.45, 0.00]
LightGBM-Fixed	0.5	-39.37	[-39.65, -39.09]	-0.19	[-0.56, 0.17]
LightGBM-Fixed	0.7	-23.28	[-23.44, -23.11]	-0.23	[-0.41, -0.04]
LightGBM-Fixed	0.9	-7.30	[-7.47, -7.14]	-0.23	[-0.42, -0.04]
XGBoost	0.1	258.93	[257.42, 260.45]	-0.72	[-0.83, -0.61]
XGBoost	0.3	259.64	[257.48, 261.83]	-0.65	[-0.97, -0.33]
XGBoost	0.5	265.45	[264.20, 266.71]	-0.39	[-0.55, -0.23]
XGBoost	0.7	272.84	[272.23, 273.45]	-0.53	[-0.68, -0.37]
XGBoost	0.9	280.63	[278.89, 282.37]	-0.38	[-0.55, -0.21]
CatBoost	0.1	-31.07	[-31.17, -30.98]	0.81	[0.62, 1.01]
CatBoost	0.3	54.47	[54.06, 54.88]	0.59	[0.47, 0.72]
CatBoost	0.5	112.62	[111.93, 113.31]	0.56	[0.35, 0.77]
CatBoost	0.7	139.84	[139.51, 140.17]	0.55	[0.37, 0.73]
CatBoost	0.9	140.26	[139.85, 140.66]	0.52	[0.35, 0.68]
Mean ablation	0.1	-84.92	[-84.96, -84.87]	1.03	[0.77, 1.29]
Mean ablation	0.3	-65.28	[-65.36, -65.20]	-1.32	[-1.54, -1.09]
Mean ablation	0.5	-46.65	[-46.74, -46.57]	-0.83	[-0.93, -0.73]
Mean ablation	0.7	-27.69	[-27.96, -27.42]	-0.52	[-0.82, -0.22]
Mean ablation	0.9	-8.85	[-9.20, -8.50]	-0.31	[-0.58, -0.05]
Exposure ablation	0.1	-84.53	[-84.57, -84.49]	1.77	[1.51, 2.03]
Exposure ablation	0.3	-65.18	[-65.26, -65.09]	0.02	[-0.11, 0.15]
Exposure ablation	0.5	-46.34	[-46.50, -46.18]	-0.04	[-0.25, 0.17]
Exposure ablation	0.7	-27.68	[-27.85, -27.50]	-0.23	[-0.42, -0.04]
Exposure ablation	0.9	-9.09	[-9.30, -8.88]	-0.23	[-0.63, 0.17]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
Bosch					
Exogenous CRAFT-GBM	0.1	-71.30	[-71.55, -71.05]	0.71	[0.66, 0.77]
Exogenous CRAFT-GBM	0.3	-52.91	[-53.08, -52.74]	0.15	[0.06, 0.24]
Exogenous CRAFT-GBM	0.5	-32.52	[-33.00, -32.05]	0.09	[0.03, 0.15]
Exogenous CRAFT-GBM	0.7	-16.38	[-16.89, -15.87]	-0.01	[-0.11, 0.10]
Exogenous CRAFT-GBM	0.9	-5.44	[-5.78, -5.10]	0.01	[-0.08, 0.10]
LightGBM-Fixed	0.1	-23.97	[-24.32, -23.61]	0.40	[0.34, 0.46]
LightGBM-Fixed	0.3	-17.17	[-17.64, -16.69]	0.17	[0.04, 0.31]
LightGBM-Fixed	0.5	-10.35	[-10.82, -9.88]	0.02	[-0.05, 0.09]
LightGBM-Fixed	0.7	-4.74	[-4.99, -4.49]	0.00	[-0.10, 0.09]
LightGBM-Fixed	0.9	-1.87	[-2.20, -1.54]	-0.03	[-0.15, 0.09]
XGBoost	0.1	-36.86	[-37.24, -36.47]	0.65	[0.55, 0.74]
XGBoost	0.3	-35.31	[-35.88, -34.75]	0.31	[0.25, 0.38]
XGBoost	0.5	-32.76	[-32.98, -32.53]	0.28	[0.17, 0.38]
XGBoost	0.7	-30.33	[-30.62, -30.04]	0.26	[0.22, 0.31]
XGBoost	0.9	-27.82	[-27.99, -27.65]	0.27	[0.20, 0.35]
CatBoost	0.1	21.43	[20.71, 22.15]	2.16	[2.09, 2.23]
CatBoost	0.3	163.80	[161.65, 165.97]	1.93	[1.86, 2.00]
CatBoost	0.5	247.92	[246.57, 249.27]	1.91	[1.85, 1.97]
CatBoost	0.7	265.57	[264.69, 266.45]	1.91	[1.84, 1.99]
CatBoost	0.9	221.57	[219.89, 223.27]	1.92	[1.84, 1.99]
Mean ablation	0.1	-71.61	[-71.78, -71.44]	1.24	[1.17, 1.31]
Mean ablation	0.3	-49.49	[-49.56, -49.41]	0.11	[0.03, 0.19]
Mean ablation	0.5	-32.75	[-32.94, -32.57]	-0.04	[-0.14, 0.07]
Mean ablation	0.7	-17.07	[-17.28, -16.86]	-0.03	[-0.14, 0.07]
Mean ablation	0.9	-6.10	[-6.37, -5.83]	0.00	[-0.11, 0.11]
Exposure ablation	0.1	-74.16	[-74.21, -74.12]	0.70	[0.64, 0.76]
Exposure ablation	0.3	-55.43	[-55.53, -55.32]	0.29	[0.23, 0.35]
Exposure ablation	0.5	-35.64	[-35.88, -35.39]	0.39	[0.35, 0.44]
Exposure ablation	0.7	-19.78	[-19.89, -19.67]	-0.07	[-0.11, -0.02]
Exposure ablation	0.9	-7.25	[-7.57, -6.92]	-0.02	[-0.10, 0.06]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
YearMSD					
Exogenous CRAFT-GBM	0.1	-44.39	[-45.15, -43.62]	1.85	[1.69, 2.01]
Exogenous CRAFT-GBM	0.3	-34.87	[-35.38, -34.37]	0.72	[0.58, 0.87]
Exogenous CRAFT-GBM	0.5	-25.78	[-26.41, -25.13]	0.56	[0.43, 0.69]
Exogenous CRAFT-GBM	0.7	-15.42	[-16.15, -14.68]	-0.03	[-0.20, 0.14]
Exogenous CRAFT-GBM	0.9	-2.74	[-3.73, -1.74]	0.02	[-0.09, 0.13]
LightGBM-Fixed	0.1	-34.30	[-34.55, -34.06]	0.29	[0.15, 0.43]
LightGBM-Fixed	0.3	-26.07	[-26.60, -25.52]	-0.07	[-0.16, 0.01]
LightGBM-Fixed	0.5	-19.18	[-19.96, -18.39]	-0.11	[-0.22, -0.01]
LightGBM-Fixed	0.7	-10.88	[-11.39, -10.37]	-0.07	[-0.15, 0.01]
LightGBM-Fixed	0.9	-1.83	[-2.47, -1.19]	0.04	[-0.05, 0.13]
XGBoost	0.1	54.21	[51.19, 57.30]	-0.21	[-0.28, -0.14]
XGBoost	0.3	47.12	[46.08, 48.16]	-0.53	[-0.64, -0.41]
XGBoost	0.5	48.75	[47.44, 50.07]	-0.54	[-0.61, -0.46]
XGBoost	0.7	53.01	[52.29, 53.74]	-0.39	[-0.42, -0.36]
XGBoost	0.9	55.65	[54.80, 56.50]	-0.43	[-0.53, -0.34]
CatBoost	0.1	27.26	[26.64, 27.89]	-0.09	[-0.18, 0.01]
CatBoost	0.3	80.31	[78.76, 81.88]	-0.26	[-0.37, -0.14]
CatBoost	0.5	112.91	[109.54, 116.34]	-0.31	[-0.41, -0.21]
CatBoost	0.7	130.27	[126.21, 134.40]	-0.31	[-0.40, -0.22]
CatBoost	0.9	123.77	[121.75, 125.82]	-0.31	[-0.37, -0.26]
Mean ablation	0.1	-46.31	[-47.24, -45.37]	5.71	[5.46, 5.96]
Mean ablation	0.3	-36.61	[-37.17, -36.04]	4.17	[3.85, 4.49]
Mean ablation	0.5	-26.16	[-27.16, -25.13]	2.81	[2.56, 3.06]
Mean ablation	0.7	-16.01	[-17.09, -14.92]	2.17	[2.06, 2.28]
Mean ablation	0.9	-2.74	[-3.14, -2.33]	1.61	[1.43, 1.80]
Exposure ablation	0.1	-44.51	[-44.79, -44.24]	1.78	[1.69, 1.87]
Exposure ablation	0.3	-34.32	[-34.89, -33.74]	0.43	[0.36, 0.50]
Exposure ablation	0.5	-24.91	[-25.34, -24.48]	0.25	[0.15, 0.35]
Exposure ablation	0.7	-14.70	[-15.42, -13.98]	0.02	[-0.09, 0.14]
Exposure ablation	0.9	-2.59	[-3.51, -1.67]	0.00	[-0.16, 0.17]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
APS Failure					
Exogenous CRAFT-GBM	0.1	-40.84	[-41.94, -39.72]	-1.39	[-2.52, -0.25]
Exogenous CRAFT-GBM	0.3	-30.45	[-32.24, -28.61]	-3.61	[-4.94, -2.26]
Exogenous CRAFT-GBM	0.5	-22.44	[-23.14, -21.74]	-3.10	[-4.16, -2.03]
Exogenous CRAFT-GBM	0.7	-11.51	[-13.10, -9.90]	-4.25	[-5.66, -2.83]
Exogenous CRAFT-GBM	0.9	-0.50	[-1.62, 0.65]	-1.08	[-2.84, 0.72]
LightGBM-Fixed	0.1	-34.16	[-35.19, -33.11]	-3.02	[-4.26, -1.76]
LightGBM-Fixed	0.3	-21.84	[-22.84, -20.83]	-5.46	[-6.10, -4.81]
LightGBM-Fixed	0.5	-13.25	[-13.99, -12.51]	-4.14	[-5.78, -2.47]
LightGBM-Fixed	0.7	-6.53	[-7.48, -5.57]	-3.32	[-4.20, -2.43]
LightGBM-Fixed	0.9	-1.70	[-2.73, -0.66]	-1.69	[-3.32, -0.04]
XGBoost	0.1	28.74	[22.57, 35.21]	-3.57	[-4.33, -2.81]
XGBoost	0.3	35.03	[33.82, 36.25]	-5.94	[-7.40, -4.46]
XGBoost	0.5	46.45	[44.70, 48.22]	-4.68	[-6.74, -2.58]
XGBoost	0.7	55.84	[53.23, 58.49]	-3.59	[-5.16, -2.00]
XGBoost	0.9	66.12	[64.64, 67.62]	-2.18	[-3.98, -0.34]
CatBoost	0.1	28.22	[25.52, 30.97]	2.39	[1.01, 3.80]
CatBoost	0.3	75.19	[72.51, 77.92]	2.09	[1.10, 3.10]
CatBoost	0.5	104.71	[97.90, 111.75]	1.16	[0.09, 2.25]
CatBoost	0.7	118.88	[110.95, 127.12]	1.93	[0.65, 3.23]
CatBoost	0.9	101.14	[97.44, 104.91]	1.78	[0.98, 2.59]
Mean ablation	0.1	-47.48	[-49.66, -45.20]	28.38	[25.06, 31.78]
Mean ablation	0.3	-33.09	[-34.36, -31.80]	5.49	[3.89, 7.11]
Mean ablation	0.5	-22.64	[-23.68, -21.58]	5.04	[3.53, 6.58]
Mean ablation	0.7	-9.67	[-10.60, -8.72]	0.19	[-0.62, 1.01]
Mean ablation	0.9	-2.83	[-4.68, -0.95]	-0.79	[-1.95, 0.39]
Exposure ablation	0.1	-42.20	[-43.21, -41.16]	1.98	[1.13, 2.83]
Exposure ablation	0.3	-30.40	[-31.34, -29.45]	-4.43	[-5.56, -3.28]
Exposure ablation	0.5	-20.10	[-21.22, -18.96]	-5.19	[-5.99, -4.39]
Exposure ablation	0.7	-10.80	[-11.89, -9.71]	-4.37	[-5.35, -3.39]
Exposure ablation	0.9	-3.75	[-4.89, -2.61]	-1.96	[-3.59, -0.30]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
KDD 2009					
Exogenous CRAFT-GBM	0.1	-17.93	[-18.39, -17.47]	4.39	[4.10, 4.67]
Exogenous CRAFT-GBM	0.3	-8.21	[-11.08, -5.24]	-2.00	[-2.48, -1.51]
Exogenous CRAFT-GBM	0.5	-2.90	[-3.61, -2.18]	-1.43	[-1.69, -1.18]
Exogenous CRAFT-GBM	0.7	2.87	[2.08, 3.67]	-1.03	[-1.28, -0.78]
Exogenous CRAFT-GBM	0.9	4.72	[4.31, 5.13]	-0.23	[-0.58, 0.13]
LightGBM-Fixed	0.1	-16.43	[-25.14, -6.71]	5.03	[4.46, 5.60]
LightGBM-Fixed	0.3	-11.14	[-16.08, -5.90]	0.63	[-0.21, 1.47]
LightGBM-Fixed	0.5	-7.56	[-8.22, -6.90]	-0.45	[-1.05, 0.15]
LightGBM-Fixed	0.7	-4.08	[-5.62, -2.51]	-0.73	[-0.95, -0.50]
LightGBM-Fixed	0.9	-1.44	[-2.30, -0.57]	-0.11	[-0.40, 0.18]
XGBoost	0.1	83.46	[79.64, 87.36]	8.33	[6.85, 9.84]
XGBoost	0.3	171.25	[165.28, 177.35]	16.16	[14.81, 17.52]
XGBoost	0.5	246.89	[241.76, 252.10]	18.66	[17.71, 19.61]
XGBoost	0.7	296.13	[288.04, 304.39]	18.86	[18.08, 19.64]
XGBoost	0.9	313.98	[309.40, 318.61]	18.75	[17.67, 19.84]
CatBoost	0.1	112.42	[100.72, 124.81]	14.11	[13.69, 14.53]
CatBoost	0.3	129.68	[111.75, 149.14]	11.77	[10.87, 12.68]
CatBoost	0.5	142.41	[135.45, 149.59]	10.46	[9.57, 11.36]
CatBoost	0.7	164.94	[136.73, 196.51]	9.81	[8.54, 11.10]
CatBoost	0.9	169.38	[150.95, 189.16]	9.67	[9.13, 10.22]
Mean ablation	0.1	-17.99	[-18.50, -17.46]	-0.66	[-0.98, -0.33]
Mean ablation	0.3	-10.76	[-12.28, -9.22]	-0.40	[-0.78, -0.01]
Mean ablation	0.5	-6.93	[-7.47, -6.38]	0.23	[-0.31, 0.78]
Mean ablation	0.7	-1.79	[-2.19, -1.39]	-0.03	[-0.42, 0.36]
Mean ablation	0.9	-0.75	[-1.59, 0.10]	-0.09	[-0.38, 0.20]
Exposure ablation	0.1	-22.93	[-23.57, -22.28]	4.39	[4.10, 4.67]
Exposure ablation	0.3	-15.87	[-16.46, -15.29]	-0.37	[-0.70, -0.03]
Exposure ablation	0.5	-10.04	[-10.98, -9.09]	-0.94	[-1.28, -0.60]
Exposure ablation	0.7	-4.67	[-5.27, -4.07]	-1.65	[-1.90, -1.39]
Exposure ablation	0.9	-1.96	[-2.45, -1.47]	-0.42	[-0.73, -0.12]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
Gisette					
Exogenous CRAFT-GBM	0.1	-78.06	[-78.26, -77.85]	-12.00	[-13.40, -10.57]
Exogenous CRAFT-GBM	0.3	-52.00	[-52.41, -51.59]	-8.56	[-10.35, -6.73]
Exogenous CRAFT-GBM	0.5	-15.97	[-16.46, -15.47]	-4.46	[-5.94, -2.97]
Exogenous CRAFT-GBM	0.7	-1.68	[-2.32, -1.04]	-3.03	[-4.22, -1.83]
Exogenous CRAFT-GBM	0.9	1.95	[1.58, 2.33]	-0.99	[-1.73, -0.23]
LightGBM-Fixed	0.1	-71.13	[-71.35, -70.90]	-13.71	[-16.24, -11.11]
LightGBM-Fixed	0.3	-51.16	[-51.42, -50.89]	-10.40	[-10.82, -9.97]
LightGBM-Fixed	0.5	-34.68	[-35.12, -34.24]	-7.66	[-9.27, -6.04]
LightGBM-Fixed	0.7	-4.28	[-5.18, -3.38]	-4.68	[-5.71, -3.64]
LightGBM-Fixed	0.9	-1.13	[-1.57, -0.69]	-2.50	[-3.82, -1.15]
XGBoost	0.1	57.19	[55.98, 58.41]	-11.85	[-12.59, -11.11]
XGBoost	0.3	70.28	[69.09, 71.48]	-10.26	[-11.26, -9.25]
XGBoost	0.5	83.61	[81.23, 86.02]	-6.82	[-8.25, -5.37]
XGBoost	0.7	96.65	[95.01, 98.30]	-3.45	[-4.49, -2.39]
XGBoost	0.9	110.16	[108.05, 112.31]	2.79	[0.22, 5.42]
CatBoost	0.1	-2.70	[-3.46, -1.93]	3.92	[3.07, 4.77]
CatBoost	0.3	136.38	[128.50, 144.53]	3.92	[1.86, 6.03]
CatBoost	0.5	282.42	[280.38, 284.48]	3.55	[2.18, 4.93]
CatBoost	0.7	326.53	[323.09, 330.00]	3.46	[1.25, 5.72]
CatBoost	0.9	316.98	[314.42, 319.55]	2.36	[1.14, 3.59]
Mean ablation	0.1	-76.20	[-76.32, -76.08]	-12.32	[-14.36, -10.24]
Mean ablation	0.3	-52.64	[-52.85, -52.43]	-6.31	[-7.79, -4.81]
Mean ablation	0.5	-34.86	[-35.17, -34.54]	-2.90	[-5.35, -0.38]
Mean ablation	0.7	-3.83	[-4.37, -3.27]	-1.71	[-3.45, 0.07]
Mean ablation	0.9	0.42	[0.09, 0.75]	-0.14	[-2.04, 1.79]
Exposure ablation	0.1	-83.84	[-83.97, -83.71]	-7.90	[-9.33, -6.45]
Exposure ablation	0.3	-65.00	[-65.28, -64.72]	-9.45	[-10.40, -8.49]
Exposure ablation	0.5	-46.80	[-47.09, -46.50]	-6.62	[-7.82, -5.40]
Exposure ablation	0.7	-6.38	[-6.55, -6.22]	-5.17	[-6.26, -4.07]
Exposure ablation	0.9	-1.80	[-2.04, -1.56]	-2.03	[-4.25, 0.24]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
Jannis					
Exogenous CRAFT-GBM	0.1	-31.20	[-31.89, -30.51]	8.00	[7.79, 8.21]
Exogenous CRAFT-GBM	0.3	-22.35	[-23.86, -20.81]	-2.03	[-2.28, -1.78]
Exogenous CRAFT-GBM	0.5	-16.07	[-18.16, -13.91]	-2.81	[-2.94, -2.69]
Exogenous CRAFT-GBM	0.7	-7.82	[-10.50, -5.07]	-1.20	[-1.28, -1.12]
Exogenous CRAFT-GBM	0.9	-2.19	[-3.51, -0.84]	-0.24	[-0.30, -0.19]
LightGBM-Fixed	0.1	-29.81	[-32.16, -27.39]	8.43	[7.94, 8.91]
LightGBM-Fixed	0.3	-19.42	[-21.95, -16.81]	1.13	[0.99, 1.28]
LightGBM-Fixed	0.5	-12.46	[-14.13, -10.76]	0.32	[0.10, 0.54]
LightGBM-Fixed	0.7	-7.09	[-11.14, -2.85]	0.11	[-0.02, 0.25]
LightGBM-Fixed	0.9	-1.69	[-5.10, 1.84]	-0.04	[-0.18, 0.10]
XGBoost	0.1	33.16	[30.22, 36.18]	9.16	[8.96, 9.36]
XGBoost	0.3	40.89	[37.86, 43.98]	0.43	[0.32, 0.55]
XGBoost	0.5	46.52	[43.44, 49.67]	-0.45	[-0.55, -0.35]
XGBoost	0.7	53.21	[49.74, 56.76]	-0.51	[-0.64, -0.37]
XGBoost	0.9	58.63	[55.80, 61.51]	-0.37	[-0.49, -0.26]
CatBoost	0.1	2.20	[-1.18, 5.70]	5.46	[5.32, 5.60]
CatBoost	0.3	26.32	[23.29, 29.42]	3.64	[3.52, 3.76]
CatBoost	0.5	49.23	[44.13, 54.52]	3.28	[3.14, 3.43]
CatBoost	0.7	59.60	[55.70, 63.60]	3.16	[2.98, 3.33]
CatBoost	0.9	62.37	[56.55, 68.40]	3.08	[2.94, 3.22]
Mean ablation	0.1	-33.57	[-34.39, -32.73]	7.76	[7.16, 8.36]
Mean ablation	0.3	-27.79	[-30.62, -24.85]	1.87	[1.74, 1.99]
Mean ablation	0.5	-19.74	[-20.95, -18.52]	-2.96	[-3.17, -2.76]
Mean ablation	0.7	-9.93	[-12.66, -7.12]	-1.25	[-1.41, -1.10]
Mean ablation	0.9	-5.57	[-8.16, -2.92]	-0.26	[-0.39, -0.13]
Exposure ablation	0.1	-34.12	[-35.55, -32.65]	5.19	[5.01, 5.36]
Exposure ablation	0.3	-25.92	[-29.61, -22.04]	1.09	[0.86, 1.31]
Exposure ablation	0.5	-16.61	[-18.00, -15.21]	0.18	[0.09, 0.27]
Exposure ablation	0.7	-11.23	[-12.39, -10.05]	-0.10	[-0.27, 0.08]
Exposure ablation	0.9	-5.32	[-8.14, -2.41]	-0.02	[-0.14, 0.10]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
HAR					
Exogenous CRAFT-GBM	0.1	-47.99	[-49.65, -46.27]	-5.74	[-10.48, -0.76]
Exogenous CRAFT-GBM	0.3	-37.44	[-37.93, -36.95]	2.91	[-0.12, 6.02]
Exogenous CRAFT-GBM	0.5	-26.82	[-27.42, -26.21]	10.81	[7.53, 14.18]
Exogenous CRAFT-GBM	0.7	-14.18	[-15.74, -12.58]	6.99	[2.87, 11.28]
Exogenous CRAFT-GBM	0.9	-0.43	[-2.09, 1.25]	-0.94	[-4.78, 3.05]
LightGBM-Fixed	0.1	-43.98	[-45.63, -42.27]	-13.62	[-24.08, -1.72]
LightGBM-Fixed	0.3	-32.76	[-33.67, -31.84]	-11.84	[-14.83, -8.74]
LightGBM-Fixed	0.5	-21.10	[-23.13, -19.02]	-5.28	[-10.37, 0.09]
LightGBM-Fixed	0.7	-12.14	[-13.33, -10.94]	-2.08	[-5.37, 1.33]
LightGBM-Fixed	0.9	-3.68	[-5.62, -1.69]	-2.57	[-5.61, 0.57]
XGBoost	0.1	198.41	[192.85, 204.08]	-15.91	[-21.70, -9.70]
XGBoost	0.3	187.78	[183.25, 192.38]	-5.15	[-10.87, 0.93]
XGBoost	0.5	193.20	[187.94, 198.55]	2.84	[-2.03, 7.96]
XGBoost	0.7	201.45	[196.81, 206.16]	5.73	[1.27, 10.39]
XGBoost	0.9	212.34	[208.75, 215.97]	6.48	[2.64, 10.47]
CatBoost	0.1	281.78	[273.56, 290.19]	0.76	[-3.29, 4.98]
CatBoost	0.3	787.10	[771.66, 802.80]	6.96	[4.20, 9.79]
CatBoost	0.5	1165.74	[1144.36, 1187.49]	9.17	[4.00, 14.61]
CatBoost	0.7	1310.60	[1284.09, 1337.62]	9.62	[5.21, 14.20]
CatBoost	0.9	1298.89	[1273.10, 1325.17]	10.87	[5.04, 17.03]
Mean ablation	0.1	-52.76	[-53.64, -51.87]	222.42	[216.23, 228.74]
Mean ablation	0.3	-32.27	[-33.27, -31.26]	139.77	[134.71, 144.94]
Mean ablation	0.5	-16.20	[-17.03, -15.37]	72.13	[69.12, 75.19]
Mean ablation	0.7	-16.60	[-17.86, -15.33]	65.11	[61.04, 69.29]
Mean ablation	0.9	-3.64	[-4.42, -2.85]	4.81	[2.62, 7.04]
Exposure ablation	0.1	-48.62	[-49.97, -47.23]	-13.00	[-17.63, -8.10]
Exposure ablation	0.3	-33.04	[-34.08, -31.99]	-9.71	[-12.15, -7.21]
Exposure ablation	0.5	-22.16	[-23.17, -21.13]	-2.80	[-3.87, -1.71]
Exposure ablation	0.7	-11.34	[-11.73, -10.96]	-3.41	[-8.94, 2.45]
Exposure ablation	0.9	-1.39	[-2.44, -0.32]	0.09	[-2.94, 3.20]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
CT Scan — descriptive fold ranges					
Exogenous CRAFT-GBM	0.1	-65.91	[-68.21, -62.02]	2.76	[-1.38, 11.99]
Exogenous CRAFT-GBM	0.3	-48.42	[-50.57, -43.19]	-3.13	[-9.04, 6.98]
Exogenous CRAFT-GBM	0.5	-32.08	[-34.52, -26.96]	-2.52	[-7.80, 0.81]
Exogenous CRAFT-GBM	0.7	-6.06	[-7.49, -4.57]	-1.49	[-5.45, 1.32]
Exogenous CRAFT-GBM	0.9	-0.30	[-1.34, 0.77]	-0.05	[-1.42, 2.07]
LightGBM-Fixed	0.1	-54.67	[-56.60, -51.07]	-1.59	[-9.80, 8.07]
LightGBM-Fixed	0.3	-40.47	[-42.05, -36.44]	-3.77	[-11.43, 3.79]
LightGBM-Fixed	0.5	-27.73	[-29.79, -23.84]	-2.76	[-6.05, 1.81]
LightGBM-Fixed	0.7	-3.16	[-4.25, -2.06]	-1.18	[-6.77, 0.70]
LightGBM-Fixed	0.9	-1.24	[-1.78, -0.66]	-0.78	[-1.93, 0.36]
XGBoost	0.1	-1.00	[-5.27, 3.68]	-1.88	[-8.71, 5.77]
XGBoost	0.3	10.22	[5.84, 19.24]	-3.11	[-11.37, 2.48]
XGBoost	0.5	13.85	[11.24, 17.97]	-2.21	[-5.58, 2.44]
XGBoost	0.7	24.23	[21.56, 29.45]	-1.18	[-6.34, 1.62]
XGBoost	0.9	34.34	[30.73, 40.05]	0.27	[-2.68, 2.93]
CatBoost	0.1	-36.43	[-42.05, -28.27]	3.20	[-0.39, 6.36]
CatBoost	0.3	5.07	[-4.29, 13.24]	2.34	[-1.10, 4.56]
CatBoost	0.5	43.85	[33.73, 65.77]	2.42	[-0.91, 5.64]
CatBoost	0.7	60.05	[47.06, 68.04]	2.03	[-3.73, 4.67]
CatBoost	0.9	49.34	[44.46, 58.44]	2.49	[-2.93, 6.39]
Mean ablation	0.1	-67.78	[-72.34, -62.48]	22.62	[3.45, 33.19]
Mean ablation	0.3	-51.42	[-53.96, -45.64]	8.75	[-2.55, 15.44]
Mean ablation	0.5	-33.55	[-36.51, -28.90]	3.28	[-1.89, 7.62]
Mean ablation	0.7	-7.35	[-9.23, -6.17]	-1.43	[-3.99, 1.19]
Mean ablation	0.9	-0.60	[-4.04, 9.41]	0.93	[-0.21, 1.86]
Exposure ablation	0.1	-67.83	[-70.68, -63.77]	6.00	[0.03, 12.63]
Exposure ablation	0.3	-52.89	[-55.20, -47.76]	-0.58	[-6.39, 8.00]
Exposure ablation	0.5	-36.69	[-38.38, -32.99]	-0.24	[-9.52, 7.53]
Exposure ablation	0.7	-8.83	[-9.32, -8.20]	-0.10	[-3.21, 2.05]
Exposure ablation	0.9	-2.88	[-3.25, -2.49]	-0.96	[-2.67, 2.62]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
Sparse 1					
Exogenous CRAFT-GBM	0.1	-81.57	[-81.88, -81.25]	-1.58	[-1.88, -1.28]
Exogenous CRAFT-GBM	0.3	-64.20	[-64.75, -63.65]	-0.74	[-0.98, -0.50]
Exogenous CRAFT-GBM	0.5	-46.36	[-46.58, -46.14]	-0.40	[-0.56, -0.23]
Exogenous CRAFT-GBM	0.7	-8.67	[-8.83, -8.50]	-0.25	[-0.42, -0.07]
Exogenous CRAFT-GBM	0.9	-2.05	[-2.31, -1.79]	-0.05	[-0.16, 0.06]
LightGBM-Fixed	0.1	-55.14	[-55.62, -54.66]	1.82	[1.29, 2.34]
LightGBM-Fixed	0.3	-43.80	[-44.04, -43.56]	0.28	[0.08, 0.48]
LightGBM-Fixed	0.5	-32.31	[-32.46, -32.17]	0.17	[-0.01, 0.35]
LightGBM-Fixed	0.7	-1.29	[-1.59, -1.00]	0.07	[-0.06, 0.21]
LightGBM-Fixed	0.9	-0.41	[-0.73, -0.09]	-0.03	[-0.10, 0.03]
XGBoost	0.1	74.36	[71.80, 76.95]	2.33	[1.18, 3.50]
XGBoost	0.3	83.50	[81.28, 85.74]	2.55	[1.43, 3.68]
XGBoost	0.5	93.48	[90.99, 96.01]	2.73	[1.55, 3.93]
XGBoost	0.7	103.93	[101.97, 105.90]	2.86	[1.65, 4.09]
XGBoost	0.9	114.86	[112.73, 117.01]	2.94	[1.73, 4.16]
CatBoost	0.1	-5.79	[-6.97, -4.60]	2.05	[0.83, 3.28]
CatBoost	0.3	106.16	[104.79, 107.52]	2.08	[0.89, 3.29]
CatBoost	0.5	182.09	[179.66, 184.53]	2.13	[0.96, 3.32]
CatBoost	0.7	212.79	[210.86, 214.72]	2.15	[0.96, 3.35]
CatBoost	0.9	212.96	[211.32, 214.61]	2.18	[1.04, 3.34]
Mean ablation	0.1	-80.87	[-81.42, -80.32]	-1.74	[-2.07, -1.41]
Mean ablation	0.3	-64.16	[-64.61, -63.71]	-0.76	[-0.92, -0.59]
Mean ablation	0.5	-46.91	[-47.15, -46.67]	-0.47	[-0.65, -0.28]
Mean ablation	0.7	-9.05	[-9.32, -8.78]	-0.18	[-0.26, -0.10]
Mean ablation	0.9	-2.57	[-2.73, -2.40]	-0.08	[-0.17, 0.01]
Exposure ablation	0.1	-84.79	[-85.46, -84.09]	1.63	[1.13, 2.12]
Exposure ablation	0.3	-68.05	[-68.53, -67.56]	0.38	[0.18, 0.59]
Exposure ablation	0.5	-50.46	[-50.88, -50.04]	0.19	[0.03, 0.35]
Exposure ablation	0.7	-10.17	[-10.39, -9.94]	0.03	[-0.10, 0.15]
Exposure ablation	0.9	-3.63	[-3.79, -3.47]	-0.02	[-0.17, 0.14]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
Sparse 2					
Exogenous CRAFT-GBM	0.1	-7.69	[-11.96, -3.22]	-0.44	[-0.66, -0.23]
Exogenous CRAFT-GBM	0.3	3.68	[1.61, 5.78]	-0.22	[-0.31, -0.14]
Exogenous CRAFT-GBM	0.5	6.27	[3.50, 9.11]	-0.10	[-0.17, -0.04]
Exogenous CRAFT-GBM	0.7	8.20	[5.59, 10.87]	-0.05	[-0.12, 0.01]
Exogenous CRAFT-GBM	0.9	8.00	[5.84, 10.20]	-0.05	[-0.10, 0.01]
LightGBM-Fixed	0.1	-6.22	[-9.02, -3.33]	2.10	[1.69, 2.51]
LightGBM-Fixed	0.3	-1.77	[-3.91, 0.41]	0.39	[0.12, 0.66]
LightGBM-Fixed	0.5	-0.86	[-1.75, 0.05]	0.18	[-0.04, 0.40]
LightGBM-Fixed	0.7	0.90	[-1.72, 3.59]	0.08	[-0.03, 0.18]
LightGBM-Fixed	0.9	0.79	[-1.26, 2.89]	-0.01	[-0.10, 0.08]
XGBoost	0.1	688.13	[667.53, 709.28]	1.04	[0.51, 1.57]
XGBoost	0.3	701.49	[675.64, 728.20]	0.72	[0.38, 1.07]
XGBoost	0.5	693.64	[686.75, 700.60]	0.42	[0.16, 0.69]
XGBoost	0.7	699.93	[687.45, 712.60]	0.28	[-0.01, 0.58]
XGBoost	0.9	712.60	[677.08, 749.74]	0.33	[0.03, 0.63]
CatBoost	0.1	532.81	[487.78, 581.30]	0.25	[-0.12, 0.62]
CatBoost	0.3	560.12	[509.30, 615.18]	0.08	[-0.21, 0.38]
CatBoost	0.5	542.24	[493.57, 594.90]	0.03	[-0.25, 0.31]
CatBoost	0.7	559.08	[518.93, 601.84]	0.06	[-0.21, 0.32]
CatBoost	0.9	544.71	[495.51, 597.97]	0.07	[-0.23, 0.36]
Mean ablation	0.1	-12.25	[-16.58, -7.70]	-0.50	[-0.74, -0.26]
Mean ablation	0.3	-0.41	[-2.13, 1.35]	-0.23	[-0.31, -0.15]
Mean ablation	0.5	-0.32	[-1.49, 0.86]	-0.12	[-0.20, -0.04]
Mean ablation	0.7	1.97	[0.14, 3.84]	-0.03	[-0.11, 0.04]
Mean ablation	0.9	4.27	[0.47, 8.22]	-0.02	[-0.12, 0.08]
Exposure ablation	0.1	-8.34	[-10.51, -6.11]	2.35	[1.60, 3.10]
Exposure ablation	0.3	-1.46	[-3.83, 0.96]	0.65	[0.33, 0.96]
Exposure ablation	0.5	-0.49	[-2.19, 1.24]	0.40	[0.16, 0.65]
Exposure ablation	0.7	1.49	[-0.49, 3.51]	0.10	[0.03, 0.17]
Exposure ablation	0.9	3.21	[2.14, 4.28]	-0.01	[-0.08, 0.05]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
Dense 1					
Exogenous CRAFT-GBM	0.1	-85.35	[-85.72, -84.97]	-0.41	[-0.60, -0.21]
Exogenous CRAFT-GBM	0.3	-67.08	[-67.56, -66.59]	-0.76	[-0.99, -0.53]
Exogenous CRAFT-GBM	0.5	-46.80	[-47.23, -46.38]	-0.55	[-0.67, -0.43]
Exogenous CRAFT-GBM	0.7	-26.53	[-27.26, -25.79]	-0.25	[-0.40, -0.09]
Exogenous CRAFT-GBM	0.9	-6.89	[-7.80, -5.97]	-0.14	[-0.27, 0.00]
LightGBM-Fixed	0.1	-77.66	[-78.23, -77.07]	97.08	[69.43, 129.24]
LightGBM-Fixed	0.3	-60.27	[-60.72, -59.81]	7.74	[5.86, 9.66]
LightGBM-Fixed	0.5	-41.38	[-42.07, -40.68]	0.66	[-0.09, 1.41]
LightGBM-Fixed	0.7	-23.29	[-23.93, -22.66]	-0.55	[-0.85, -0.24]
LightGBM-Fixed	0.9	-7.90	[-8.56, -7.23]	-0.38	[-0.56, -0.19]
XGBoost	0.1	191.37	[184.16, 198.76]	69.23	[53.78, 86.24]
XGBoost	0.3	225.26	[223.11, 227.44]	7.68	[4.75, 10.69]
XGBoost	0.5	233.10	[231.35, 234.87]	0.04	[-0.59, 0.68]
XGBoost	0.7	241.22	[238.54, 243.92]	-0.73	[-0.91, -0.54]
XGBoost	0.9	251.12	[246.81, 255.49]	-0.54	[-0.73, -0.35]
CatBoost	0.1	-24.55	[-26.46, -22.58]	24.37	[17.24, 31.93]
CatBoost	0.3	75.43	[70.43, 80.59]	-1.40	[-1.99, -0.80]
CatBoost	0.5	151.08	[145.61, 156.67]	-2.81	[-3.17, -2.45]
CatBoost	0.7	188.32	[181.71, 195.10]	-3.09	[-3.40, -2.79]
CatBoost	0.9	193.84	[186.37, 201.49]	-3.20	[-3.50, -2.89]
Mean ablation	0.1	-85.78	[-86.05, -85.51]	-0.63	[-0.81, -0.45]
Mean ablation	0.3	-68.20	[-68.51, -67.89]	-0.32	[-0.48, -0.16]
Mean ablation	0.5	-48.48	[-48.89, -48.07]	-0.20	[-0.32, -0.08]
Mean ablation	0.7	-28.78	[-29.74, -27.80]	-0.08	[-0.25, 0.10]
Mean ablation	0.9	-9.36	[-10.01, -8.71]	-0.06	[-0.23, 0.12]
Exposure ablation	0.1	-85.34	[-85.55, -85.14]	80.26	[71.92, 89.00]
Exposure ablation	0.3	-67.63	[-68.23, -67.02]	60.91	[60.24, 61.59]
Exposure ablation	0.5	-47.99	[-48.96, -47.01]	45.61	[10.84, 91.29]
Exposure ablation	0.7	-28.50	[-29.20, -27.79]	1.02	[-0.02, 2.08]
Exposure ablation	0.9	-10.62	[-11.38, -9.86]	-0.19	[-0.34, -0.03]

Continued on next page

Table 11 (continued)

Model	f	Δ time (%)	Interval	Δ loss (%)	Interval
Dense 2					
Exogenous CRAFT-GBM	0.1	-84.47	[-84.81, -84.12]	-2.13	[-2.58, -1.68]
Exogenous CRAFT-GBM	0.3	-66.20	[-66.62, -65.78]	-1.03	[-1.17, -0.89]
Exogenous CRAFT-GBM	0.5	-46.27	[-47.02, -45.52]	-0.60	[-0.88, -0.32]
Exogenous CRAFT-GBM	0.7	-26.05	[-26.89, -25.20]	-0.29	[-0.43, -0.14]
Exogenous CRAFT-GBM	0.9	-6.97	[-7.78, -6.15]	-0.10	[-0.15, -0.06]
LightGBM-Fixed	0.1	-75.22	[-75.93, -74.48]	5.49	[3.17, 7.87]
LightGBM-Fixed	0.3	-59.55	[-60.04, -59.07]	1.05	[0.46, 1.64]
LightGBM-Fixed	0.5	-41.43	[-41.90, -40.95]	0.24	[-0.11, 0.58]
LightGBM-Fixed	0.7	-23.74	[-24.23, -23.25]	0.12	[0.03, 0.21]
LightGBM-Fixed	0.9	-7.87	[-8.85, -6.89]	0.06	[-0.16, 0.27]
XGBoost	0.1	202.67	[195.48, 210.04]	0.70	[-0.67, 2.09]
XGBoost	0.3	214.66	[206.02, 223.53]	-2.74	[-3.14, -2.33]
XGBoost	0.5	223.12	[217.08, 229.26]	-3.38	[-4.04, -2.72]
XGBoost	0.7	238.49	[233.24, 243.82]	-3.62	[-4.42, -2.81]
XGBoost	0.9	253.17	[247.12, 259.34]	-3.62	[-4.40, -2.84]
CatBoost	0.1	-14.02	[-15.34, -12.67]	1.88	[0.84, 2.93]
CatBoost	0.3	92.24	[88.68, 95.87]	-0.46	[-1.11, 0.19]
CatBoost	0.5	173.34	[169.57, 177.16]	-0.88	[-1.65, -0.10]
CatBoost	0.7	215.85	[207.97, 223.93]	-1.10	[-1.82, -0.37]
CatBoost	0.9	224.14	[217.00, 231.44]	-1.31	[-2.16, -0.44]
Mean ablation	0.1	-85.17	[-85.63, -84.69]	-2.73	[-3.25, -2.21]
Mean ablation	0.3	-67.11	[-67.64, -66.57]	-0.97	[-1.17, -0.77]
Mean ablation	0.5	-47.68	[-48.11, -47.25]	-0.54	[-0.78, -0.29]
Mean ablation	0.7	-27.56	[-28.35, -26.76]	-0.16	[-0.42, 0.10]
Mean ablation	0.9	-8.83	[-9.32, -8.34]	-0.01	[-0.14, 0.11]
Exposure ablation	0.1	-83.74	[-84.36, -83.09]	7.43	[4.30, 10.66]
Exposure ablation	0.3	-67.16	[-67.45, -66.86]	1.41	[0.49, 2.33]
Exposure ablation	0.5	-48.70	[-48.95, -48.45]	1.03	[0.36, 1.70]
Exposure ablation	0.7	-29.55	[-30.66, -28.42]	0.12	[0.03, 0.21]
Exposure ablation	0.9	-11.24	[-11.82, -10.64]	0.03	[-0.09, 0.14]

Table 12: RQ2: synthetic oracle diagnostics used only for evaluation. Values are mean ± sample SD over five generated problems. The first metric is oracle cross-entropy for Sparse 1/2 and function MSE for Dense 1, while Bernoulli KL is defined only for Sparse 1/2. These quantities do not guide training or selection. Full (CRAFT-GBM ref.) identifies the fresh Exogenous CRAFT-GBM reference. Full (ablation ref.) identifies the reference shared by Mean and Exposure ablations.

Model	f	Oracle CE / function MSE	Bernoulli KL
Sparse 1 — Oracle cross-entropy and Bernoulli KL (nats/observation)			
LightGBM-Full (CRAFT-GBM ref.)	—	0.4969 ± 0.0011	0.0245 ± 0.0010
LightGBM-Full (ablation ref.)	—	0.4969 ± 0.0011	0.0245 ± 0.0010
LightGBM-Full	—	0.4969 ± 0.0011	0.0245 ± 0.0010
Exogenous CRAFT-GBM	0.1	0.4889 ± 0.0014	$0.0165 \pm 8.63\times10^{-4}$
Exogenous CRAFT-GBM	0.3	0.4932 ± 0.0012	$0.0208 \pm 7.72\times10^{-4}$
Exogenous CRAFT-GBM	0.5	$0.4951 \pm 9.26\times10^{-4}$	$0.0227 \pm 7.03\times10^{-4}$
Exogenous CRAFT-GBM	0.7	$0.4959 \pm 9.45\times10^{-4}$	$0.0235 \pm 8.55\times10^{-4}$
Exogenous CRAFT-GBM	0.9	0.4968 ± 0.0011	0.0244 ± 0.0012
LightGBM-Fixed	0.1	0.5059 ± 0.0020	0.0335 ± 0.0012
LightGBM-Fixed	0.3	0.4985 ± 0.0012	$0.0261 \pm 7.38\times10^{-4}$
LightGBM-Fixed	0.5	0.4977 ± 0.0012	$0.0253 \pm 7.43\times10^{-4}$
LightGBM-Fixed	0.7	0.4973 ± 0.0012	$0.0249 \pm 8.93\times10^{-4}$
LightGBM-Fixed	0.9	0.4970 ± 0.0012	0.0246 ± 0.0013
XGBoost	0.1	0.5086 ± 0.0040	0.0362 ± 0.0032
XGBoost	0.3	0.5097 ± 0.0040	0.0374 ± 0.0032
XGBoost	0.5	0.5106 ± 0.0041	0.0382 ± 0.0033
XGBoost	0.7	0.5112 ± 0.0041	0.0389 ± 0.0033
XGBoost	0.9	0.5116 ± 0.0042	0.0393 ± 0.0034
CatBoost	0.1	0.5073 ± 0.0044	0.0349 ± 0.0036
CatBoost	0.3	0.5076 ± 0.0042	0.0352 ± 0.0034
CatBoost	0.5	0.5078 ± 0.0041	0.0354 ± 0.0033
CatBoost	0.7	0.5079 ± 0.0041	0.0355 ± 0.0034
CatBoost	0.9	0.5080 ± 0.0042	0.0356 ± 0.0034
Mean ablation	0.1	0.4880 ± 0.0013	$0.0157 \pm 5.64\times10^{-4}$
Mean ablation	0.3	0.4931 ± 0.0011	$0.0207 \pm 8.74\times10^{-4}$
Mean ablation	0.5	0.4948 ± 0.0011	$0.0224 \pm 9.20\times10^{-4}$
Mean ablation	0.7	0.4959 ± 0.0010	$0.0235 \pm 9.94\times10^{-4}$
Mean ablation	0.9	0.4966 ± 0.0010	$0.0242 \pm 9.90\times10^{-4}$
Exposure ablation	0.1	0.5051 ± 0.0023	0.0327 ± 0.0016
Exposure ablation	0.3	0.4988 ± 0.0012	$0.0264 \pm 7.72\times10^{-4}$
Exposure ablation	0.5	0.4979 ± 0.0011	$0.0255 \pm 8.45\times10^{-4}$
Exposure ablation	0.7	0.4971 ± 0.0011	0.0248 ± 0.0010
Exposure ablation	0.9	0.4970 ± 0.0011	$0.0247 \pm 9.67\times10^{-4}$

Continued on next page

Table 12 (continued)

Model	f	Oracle CE / function MSE	Bernoulli KL
Sparse 2 — Oracle cross-entropy and Bernoulli KL (nats/observation)			
LightGBM-Full (CRAFT-GBM ref.)	—	0.4652 ± 0.0021	0.0105 ± 0.0019
LightGBM-Full (ablation ref.)	—	0.4652 ± 0.0021	0.0105 ± 0.0019
LightGBM-Full	—	0.4652 ± 0.0021	0.0105 ± 0.0019
Exogenous CRAFT-GBM	0.1	0.4628 ± 0.0022	0.0082 ± 0.0019
Exogenous CRAFT-GBM	0.3	0.4643 ± 0.0019	0.0096 ± 0.0016
Exogenous CRAFT-GBM	0.5	0.4647 ± 0.0020	0.0101 ± 0.0018
Exogenous CRAFT-GBM	0.7	0.4648 ± 0.0020	0.0102 ± 0.0017
Exogenous CRAFT-GBM	0.9	0.4651 ± 0.0021	0.0104 ± 0.0019
LightGBM-Fixed	0.1	0.4751 ± 0.0032	0.0205 ± 0.0030
LightGBM-Fixed	0.3	0.4671 ± 0.0029	0.0125 ± 0.0027
LightGBM-Fixed	0.5	0.4661 ± 0.0026	0.0115 ± 0.0024
LightGBM-Fixed	0.7	0.4655 ± 0.0024	0.0109 ± 0.0022
LightGBM-Fixed	0.9	0.4651 ± 0.0022	0.0105 ± 0.0020
XGBoost	0.1	0.4701 ± 0.0037	0.0154 ± 0.0034
XGBoost	0.3	0.4685 ± 0.0026	0.0139 ± 0.0024
XGBoost	0.5	0.4671 ± 0.0020	0.0124 ± 0.0017
XGBoost	0.7	0.4664 ± 0.0017	0.0118 ± 0.0014
XGBoost	0.9	0.4666 ± 0.0016	0.0120 ± 0.0013
CatBoost	0.1	0.4663 ± 0.0027	0.0117 ± 0.0025
CatBoost	0.3	0.4655 ± 0.0023	0.0109 ± 0.0021
CatBoost	0.5	0.4653 ± 0.0021	0.0107 ± 0.0019
CatBoost	0.7	0.4654 ± 0.0019	0.0107 ± 0.0017
CatBoost	0.9	0.4654 ± 0.0020	0.0108 ± 0.0017
Mean ablation	0.1	0.4627 ± 0.0013	$0.0081 \pm 9.63 \times 10^{-4}$
Mean ablation	0.3	0.4642 ± 0.0017	0.0096 ± 0.0015
Mean ablation	0.5	0.4646 ± 0.0019	0.0100 ± 0.0016
Mean ablation	0.7	0.4649 ± 0.0020	0.0103 ± 0.0018
Mean ablation	0.9	0.4650 ± 0.0021	0.0104 ± 0.0019
Exposure ablation	0.1	0.4762 ± 0.0043	0.0215 ± 0.0041
Exposure ablation	0.3	0.4682 ± 0.0029	0.0136 ± 0.0027
Exposure ablation	0.5	0.4670 ± 0.0027	0.0124 ± 0.0025
Exposure ablation	0.7	0.4655 ± 0.0023	0.0109 ± 0.0021
Exposure ablation	0.9	0.4652 ± 0.0022	0.0106 ± 0.0020

Continued on next page

Table 12 (continued)

Model	f	Oracle CE / function MSE	Bernoulli KL
Dense 1 — Function MSE (squared target units)			
LightGBM-Full (CRAFT-GBM ref.)	—	0.0922 ± 0.0052	—
LightGBM-Full (ablation ref.)	—	0.0922 ± 0.0052	—
LightGBM-Full	—	0.0922 ± 0.0052	—
Exogenous CRAFT-GBM	0.1	0.0833 ± 0.0053	—
Exogenous CRAFT-GBM	0.3	0.0750 ± 0.0032	—
Exogenous CRAFT-GBM	0.5	0.0797 ± 0.0045	—
Exogenous CRAFT-GBM	0.7	0.0862 ± 0.0038	—
Exogenous CRAFT-GBM	0.9	0.0893 ± 0.0029	—
LightGBM-Fixed	0.1	3.3435 ± 1.0670	—
LightGBM-Fixed	0.3	0.2683 ± 0.0323	—
LightGBM-Fixed	0.5	0.1069 ± 0.0095	—
LightGBM-Fixed	0.7	0.0803 ± 0.0042	—
LightGBM-Fixed	0.9	0.0837 ± 0.0036	—
XGBoost	0.1	2.1595 ± 0.4601	—
XGBoost	0.3	0.2679 ± 0.0537	—
XGBoost	0.5	0.0938 ± 0.0098	—
XGBoost	0.7	0.0766 ± 0.0043	—
XGBoost	0.9	0.0799 ± 0.0041	—
CatBoost	0.1	0.6961 ± 0.1549	—
CatBoost	0.3	0.0616 ± 0.0089	—
CatBoost	0.5	0.0312 ± 0.0017	—
CatBoost	0.7	$0.0250 \pm 9.57 \times 10^{-4}$	—
CatBoost	0.9	0.0234 ± 0.0014	—
Mean ablation	0.1	0.0776 ± 0.0022	—
Mean ablation	0.3	0.0852 ± 0.0043	—
Mean ablation	0.5	0.0871 ± 0.0035	—
Mean ablation	0.7	0.0901 ± 0.0027	—
Mean ablation	0.9	0.0908 ± 0.0035	—
Exposure ablation	0.1	2.5590 ± 0.2552	—
Exposure ablation	0.3	1.8285 ± 0.0059	—
Exposure ablation	0.5	1.4653 ± 0.7896	—
Exposure ablation	0.7	0.1145 ± 0.0189	—
Exposure ablation	0.9	0.0875 ± 0.0045	—

K When Can Dynamic Feature Screening Reduce Work?

Forming all bounded endpoints $L_{j,t}, U_{j,t}$ from cached boundary values $u_\rho(v^+;\alpha)$ costs $O(p)$. New boundary-cache entries require the numerical root calculation described in Appendix B. The default rule for a set of possible-best features S_t then finds $C_t = \max_j L_{j,t}$ and retains every $U_{j,t} \geq C_t$ in two linear scans, without sorting. In Exogenous K mode, the implementation applies `std::nth_element` with the total point-estimate $c_{j,t}$ ordering, then sorts the K_f retained internal positions for the downstream mask. Its expected selection cost is $O(p + K_f \log K_f)$, with $O(p)$ reusable scratch. This is not a worst-case linear-time claim. Boundary numerators $u_\rho(v^+;\alpha)$ are memoized on the upward-rounded squared-error grid v^+, shared across features, so each occupied grid point requires a root calculation only once. The current observation count $N_{j,t-1}$ remains the radius $r_{j,t-1}$ denominator.

Before each output tree, the implementation clears and collects the root-gain $G_{j,m,t}$ slots for the $K_t = |S_t|$ selected features. Collecting and transforming these gains costs $O(MK_t)$. When the selection S_t changes, the incremental mask update costs $O(K_{t-1} + K_t)$, whereas an unchanged set requires no mask or list rebuild. One-time feature initialization costs $O(p \log p)$, including the sort of canonical feature IDs. The gain histories, endpoint scratch, and top-K scratch require $O(p)$ storage, and the shared boundary cache requires $O(128[1+\log(1+V_{\max})])$ entries, where $V_{\max} = \max_j V_{j,t}$. Updating the squared-error sums $V_{j,t}$ costs $O(K_t)$ per completed round. The selected-count history additionally stores one value per completed round, requiring $O(t)$ space. The root-gain buffer and fixed feature mappings require $O(p)$, in addition to native Dataset storage. CRAFT-GBM stores no modified or rebuilt EFB representation.

Let e_a be method a's realized completed-round endpoint, $F_{a,t}$ its feature-dependent histogram and split-search work, and $m_{a,t}$ its training-loss evaluation and endpoint-controller work. Let $b_{a,t}$ include remaining native per-round work, such as gradient and Hessian calculation, row sampling, and prediction updates. For CRAFT-GBM, let $a_{a,t} \geq 0$ be mapping, interval, selection, masking, and gain-collection work, which is zero for native LightGBM. If $C_a^{(0)}$ contains dataset loading, initial-loss evaluation, model materialization, and other method-specific endpoint costs, the honest decomposition is

$$C_a = C_a^{(0)} + \sum_{t=1}^{e_a} \big(F_{a,t} + b_{a,t} + m_{a,t} + a_{a,t}\big), \qquad a_{\mathrm{LGBM},t} = 0. \tag{21}$$

In RQ1, the endogenous endpoint is $e_a = \tau_a$, whereas fixed-horizon comparators use the transferred pre-run endpoint $t^{(0)}$. At a common fixed endpoint, CRAFT-GBM is faster only when its saved histogram and split-search work $F_{a,t}$ exceeds its added selector work $a_{a,t}$ and the net change in other native, metric/controller, and endpoint costs $C_a^{(0)}$. Under adaptive stopping, tail length also enters. The measured runtime consequently includes the entire executed prefix, any patience tail, training-loss L_t evaluations, the controller, and model serialization. Final-test prediction is outside this timing scope.

For physical accounting, let $\mathcal{H}_{a,t} \subseteq \mathcal{E}$ be the EFB groups actually visited and $\mathcal{J}_{a,t} \subseteq \{1,\dots,p\}$ the features actually considered as split candidates in arm $a \in \{\mathrm{L},\mathrm{J}\}$. One may decompose feature-dependent native work as

$$F_{a,t} = \sum_{b \in \mathcal{H}_{a,t}} g^a_{b,t} + \sum_{j \in \mathcal{J}_{a,t}} e^a_{j,t},$$

where the first term includes shared packed-storage and histogram work and the second includes feature-specific split evaluation. Native LightGBM has $\mathcal{J}_{\mathrm{L},t}$ equal to its native candidate set, and to $\{1,\dots,p\}$ with all native features eligible. A conforming CRAFT-GBM implementation has

$$\mathcal{J}_{\mathrm{J},t} = S_t.$$

Its directly implicated groups, and exactly the groups invoked in the validated forced-column-wise scope, are $\mathcal{H}_{\mathrm{J},t} = \mathcal{E}(S_t)$. Unchanged native heuristics may still process bins of unselected features inside an implicated group. Physical costs differ with group membership, density, bin width, feature fraction, candidate count, memory traffic, and native sharing. Thus neither the selection size K_t nor the feature fraction K_t/p, nor even $|\mathcal{E}(S_t)|/B$, is by itself a physical-work guarantee. If $p = 1$, both models make the sole feature eligible while CRAFT-GBM still pays controller overhead, ruling out a universal no-slowdown theorem.