GLASS-BRW

Precision-Gated Rule Ensemble with Abstention

Disclaimer

This document outlines a theoretical framework and initial system design. Parameter values and constraints are subject to change as experimentation progresses; however, the overarching methodology and design philosophy will be preserved.

1 Problem Setting

- $X \in \mathbb{R}^d$ denotes customer feature vectors.
- $Y \in \{0, 1\}$ denotes the target variable (NO SUBSCRIBE, SUBSCRIBE).
- $S = \phi(X)$ denotes a deterministic segment assignment function.

$$S = \phi(X)$$

- The function $\phi(\cdot)$ maps each customer to a finite categorical lattice.
- Segment assignments are deterministic and fixed at inference time.
- GLASS-BRW does **not** learn a classifier $f : X \to Y$.
- Instead, it learns a partial decision function:

$$f_{\text{GLASS}} : X \to \{-1, 0, 1\}$$

- -1 denotes explicit **ABSTAIN**.
- Predictions are emitted only when symbolic segment rules meet strict precision guarantees.

2 Rule Representation

- A rule r is defined as an ordered pair:

$$r = (C_r, y_r)$$

- C_r is a conjunction of segment conditions.
- $y_r \in \{0, 1\}$ is the predicted class.

$$C_r = \{(f_1 = \ell_1), \ldots, (f_k = \ell_k)\}$$

- Each $(f_i = \ell_i)$ corresponds to a segment constraint.
- Rule complexity is bounded by $k \leq 3$.
- A customer x matches rule r if and only if:

$$\forall (f, \ell) \in C_r : \phi_f(x) = \ell$$

3 Candidate Rule Generation

3.1 Feature Pre-Filtering

Let:

- F denote the full set of segment features
- L_f denote the discrete levels of feature f
- allowed_features $\subseteq F$ be a pre-selected subset chosen via permutation importance or mutual information (typically top 20–40 features)

3.2 Lattice Enumeration with Quality Gates

Candidate rules are generated via bounded lattice enumeration using only features in allowed_features:

$$\mathcal{R} = \bigcup_{k=1}^{3} \left\{ r \;\middle|\; C_r \subseteq \prod_{i=1}^{k} (f_i, \ell_i),\ f_i \in \text{allowed\_features},\ |\text{supp}(r)| \geq n_{\min} \right\}$$

where $n_{\min}$ is a minimum support threshold (e.g. 50 samples).

3.3 Enhanced Quality Metrics

For each candidate rule r with predicted class y_r, we compute on held-out validation data:

Base Statistics

- True positives TP_r, false positives FP_r, true negatives TN_r, false negatives FN_r
- Precision: $p_r = \frac{\text{TP}_r}{\text{TP}_r + \text{FP}_r}$ (if $y_r = 1$) or $p_r = \frac{\text{TN}_r}{\text{TN}_r + \text{FN}_r}$ (if $y_r = 0$)
- Coverage: $c_r = \frac{|\{x : x \models r\}|}{|X|}$

Information-Theoretic Metrics

- Information gain: $\text{IG}_r = H(Y) - H(Y|r)$ where $H(Y)$ is the entropy of the validation set and $H(Y|r)$ is the conditional entropy given the rule
- Segment entropy: $H_{\text{seg}} = -p_r \log_2(p_r) - (1 - p_r) \log_2(1 - p_r)$

Business-Aligned Metrics

- Lift (for subscribe rules): $\text{lift}_r = \frac{p_r}{\text{base\_rate}_{\text{subscribe}}}$ where $\text{base\_rate}_{\text{subscribe}}$ is the proportion of subscribers in validation
- Relative recall (for subscribe rules): $\text{rr} = \frac{\text{TP}_r}{\text{total\_subscribers}}$
- Absolute recall: TP_r

For no subscribe rules, analogous metrics are computed using TN_r and the base rate of class 0.

3.4 Multi-Stage Filtering

Rules must pass all of the following gates to enter the pool $\mathcal{R}$:

Class 1 (Subscribe) Rules

1. Precision: $p_r \geq p_{\min}$ (typically 0.75)
2. Information gain: $\text{IG}_r \geq \text{min\_info\_gain}_{\text{subscribe}}$ (e.g. 0.01)
3. Entropy ceiling: $H_{\text{seg}} \leq \text{max\_entropy}_{\text{subscribe}}$ (e.g. 0.85)
4. Lift requirement: $\text{lift}_r \geq \text{min\_lift}_{\text{subscribe}}$ (e.g. 2.0)
5. Recall contribution: $\text{rr} \geq \text{min\_recall}_{\text{subscribe}}$ (e.g. 0.005) AND $\text{TP}_r \geq \text{min\_tp}_{\text{subscribe}}$ (e.g. 10)

Class 0 (No Subscribe) Rules Similar constraints apply with potentially different thresholds:

1. Precision: $p_r \geq p_{\min}$ (typically 0.75)
2. Information gain: $\text{IG}_r \geq \text{min\_info\_gain}_{\text{no\_subscribe}}$ (e.g. 0.01)
3. Entropy ceiling: $H_{\text{seg}} \leq \text{max\_entropy}_{\text{no\_subscribe}}$ (e.g. 0.85)
4. Lift and recall thresholds adapted for class 0

The resulting filtered rule pool satisfies $|\mathcal{R}| \approx 200$–$500$.

4 Rule Metrics (Validation-Based)

All metrics are computed on held-out validation data.

4.1 Precision (Primary Gate)

$$p_r = \frac{\mathrm{TP}_r}{\mathrm{TP}_r + \mathrm{FP}_r}$$

- Rules with $p_r < p_{\min}$ (typically 0.75) are ineligible for selection

4.2 Coverage

$$c_r = \frac{|\{x : x \models r\}|}{|X|}$$

- Measures the fraction of customers covered by rule r

4.3 Interpretability

$$i_r = \frac{1}{|C_r| + 1}$$

- Penalizes rule complexity
- Biases selection toward simpler symbolic explanations

4.4 Stability

$$\text{stability}_r = 1 - \frac{\text{std}(p_r^{(j)})}{\mathbb{E}[p_r^{(j)}]}$$

- $p_r^{(j)}$ denotes precision across cross-validation folds
- Low stability indicates non-stationarity or sample noise

4.5 Ensemble-Aware Metrics

Given a downstream EBM with output $\hat{p}_{\text{EBM}}(x)$:

Boundary Ambiguity

$$a_r = \mathbb{E}_{x \models r}\left[|\hat{p}_{\text{EBM}}(x) - 0.5|\right]$$

- Lower values indicate regions where EBM is uncertain

EBM Overlap

$$o_r = \mathbb{P}_{x \models r}\left(|\hat{p}_{\text{EBM}}(x) - 0.5| > 0.20\right)$$

- Higher values indicate redundancy with EBM's confident predictions

5 Rule Selection as Integer Optimization

5.1 Decision Variables

$$x_r \in \{0, 1\} \quad \forall r \in \mathcal{R}$$

5.2 Objective Function

$$\max \sum_{r \in \mathcal{R}} x_r \left(p_r^2 c_r i_r - \lambda_1 a_r - \lambda_2 o_r \right)$$

- Quadratic precision enforces superlinear preference for high-purity rules
- Penalties encode ensemble complementarity

5.3 Constraints

Cardinality

$$8 \leq \sum_r x_r \leq 10$$

Minimum Coverage

$$\sum_r x_r c_r \geq 0.60$$

Precision Gate

$$x_r = 0 \quad \text{if } p_r < p_{\min}$$

Information Gain Floor

$$x_r = 0 \quad \text{if } \mathrm{IG}_r < \mathrm{min\_info\_gain}_{y_r}$$

Entropy Ceiling

$$x_r = 0 \quad \text{if } H_{\text{seg}} > \mathrm{max\_entropy}_{y_r}$$

Lift Requirement (Subscribe Rules)

$$x_r = 0 \quad \text{if } y_r = 1 \text{ and } \mathrm{lift}_r < \mathrm{min\_lift}_{\text{subscribe}}$$

Recall Contribution (Subscribe Rules) Dual recall gates ensure rules capture meaningful volumes:

$$x_r = 0 \quad \text{if } y_r = 1 \text{ and } (\mathrm{TP}_r < \mathrm{min\_tp}_{\text{subscribe}} \text{ or } \mathrm{rr} < \mathrm{min\_recall}_{\text{subscribe}})$$

where $\mathrm{rr} = \frac{\mathrm{TP}_r}{\mathrm{total\_subscribers}}$ is relative recall.

- Absolute threshold prevents spurious rules from small folds
- Relative threshold ensures population-level impact

Class Balance

$$3 \leq \sum_r x_r \mathbb{I}(y_r = 1) \leq 5$$

$$4 \leq \sum_r x_r \mathbb{I}(y_r = 0) \leq 6$$

Feature Diversity

$$\sum_{r: f \in C_r} x_r \leq 3 \quad \forall f \in F$$

5.4 Suggested Hyperparameter Starting Points

For binary classification with base rate ≈ 0.12:

Parameter	Subscribe Rules	No Subscribe Rules
$p_{\min}$	0.75	0.75
min_info_gain	0.01	0.01
max_entropy	0.85	0.85
min_lift	2.0	2.0*
min_tp	10	10*
min_recall	0.005	0.005*
$n_{\min}$ (support)	50	50
|allowed_features|	20–40 (via permutation importance)	

*For no subscribe rules, use analogous metrics computed on TN_r and class 0 base rate. These thresholds balance rule quality with pool size, typically yielding 200–500 candidates.

6 Execution Semantics (First-Match-Wins)

Selected rules $\mathcal{R}^*$ are ordered by descending p_r. Prediction for a customer x is defined as:

$$f_{\text{GLASS}}(x) = \begin{cases} y_r, & \exists r \in \mathcal{R}^* \text{ s.t. } x \models r \\ -1, & \text{otherwise} \end{cases}$$

- Mutual exclusivity
- Determinism
- Single-rule attribution

7 Probabilistic Output Encoding

$$P(Y = 1 \mid x) = \begin{cases} p_r, & f_{\text{GLASS}}(x) = 1 \\ 1 - p_r, & f_{\text{GLASS}}(x) = 0 \\ 0.5, & f_{\text{GLASS}}(x) = -1 \end{cases}$$

- Abstention propagates uncertainty to downstream models

8 Guarantees

Theorem 1 (Precision Lower Bound)

$$\mathbb{P}(Y = \hat{Y} \mid f_{\text{GLASS}}(x) \neq -1) \geq p_{\min}$$

- Each covered instance matches exactly one rule with precision $\geq p_{\min}$

Theorem 2 (Interpretability Bound)

- Every prediction is explained by at most three human-interpretable conditions

Theorem 3 (Ensemble Complementarity)

- For $\lambda_2 > 0$, the optimizer prefers rules covering regions of low EBM confidence