A Calibration-Aware Framework for Geometric Support Evaluation of Synthetic Tabular Data

Lee Carlin Yuval Benjamini The Hebrew University of Jerusalem

Why Current Tabular Support Evaluation is Unreliable

- Evaluating synthetic generators is an ongoing challenge, especially for tabular mixed-type datasets.
- Characterized by high-cardinality features, non-linear dependencies, and lack of canonical representations.
- Practitioners have adopted *Geometric Support Metrics* (Precision/Recall) from Image/NLP domains.
- However, they assume a reliable embedding space. In tabular data, used as "black-box" with unknown validity.
- Sparse categorical structures and conditional constraints distort support estimation, leading to unreliable diagnostics.

Support Metrics: Formal Setup

Date: $\mathcal{D}_{\text{real}} = \{x_i\}_{i=1}^{N}$, $\mathcal{D}_{\text{syn}} = \{y_j\}_{j=1}^{M} \subset \mathcal{X}$.
Embedder: $\Phi : \mathcal{X} \to \mathbb{R}^p$ maps $\mathcal{Z}_{\text{real}} = \{\Phi(x_i)\}_{i=1}^{N} \sim P$, $\mathcal{Z}_{\text{syn}} = \{\Phi(y_j)\}_{j=1}^{M} \sim Q$
The support $S_P(\alpha)$ denotes a high-density region of P, for a probability level α:

$$S_P(\alpha) = \{z \in \mathbb{R}^p : f(z) \geq \tau_\alpha\}.$$

Precision (Fidelity):

$$\text{Pr}(P, Q) = 1 - \Pr_{z \sim Q}\left[z \in \widehat{S}_P(\alpha)\right]$$

Fraction of synthetic samples *inside* real support.

Recall (Diversity):

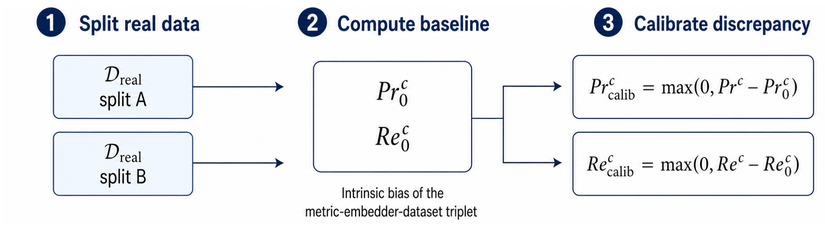
$$\text{Re}(P, Q) = 1 - \Pr_{z \sim P}\left[z \in \widehat{S}_Q(\alpha)\right]$$

Fraction of real samples *covered* by synthetic support.

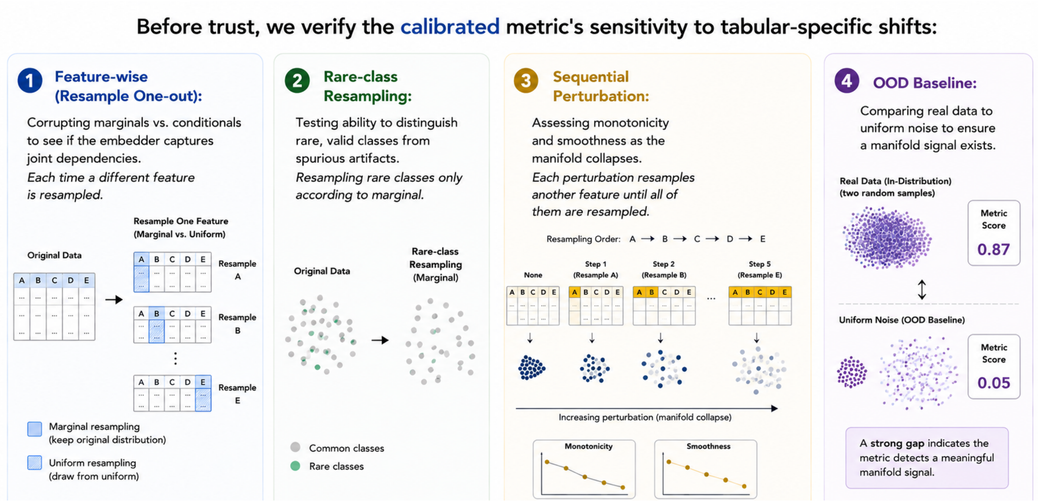
Baseline Bias

- **Empirical Bias:** Support metrics often assign non-zero discrepancy to samples drawn from *identical* distributions.
- **Sensitivity:** Raw scores (Pr, Re) are highly sensitive to:
  1. Embedding architecture (PCA vs. DSVDD vs. Contrastive).
  2. Preprocessing (Scaling, Encoding).
  3. Hyperparameters (e.g., k-nearest neighbors).
- **Inconsistency:** Generators are often ranked differently depending purely on the choice of the embedder, not the data quality.

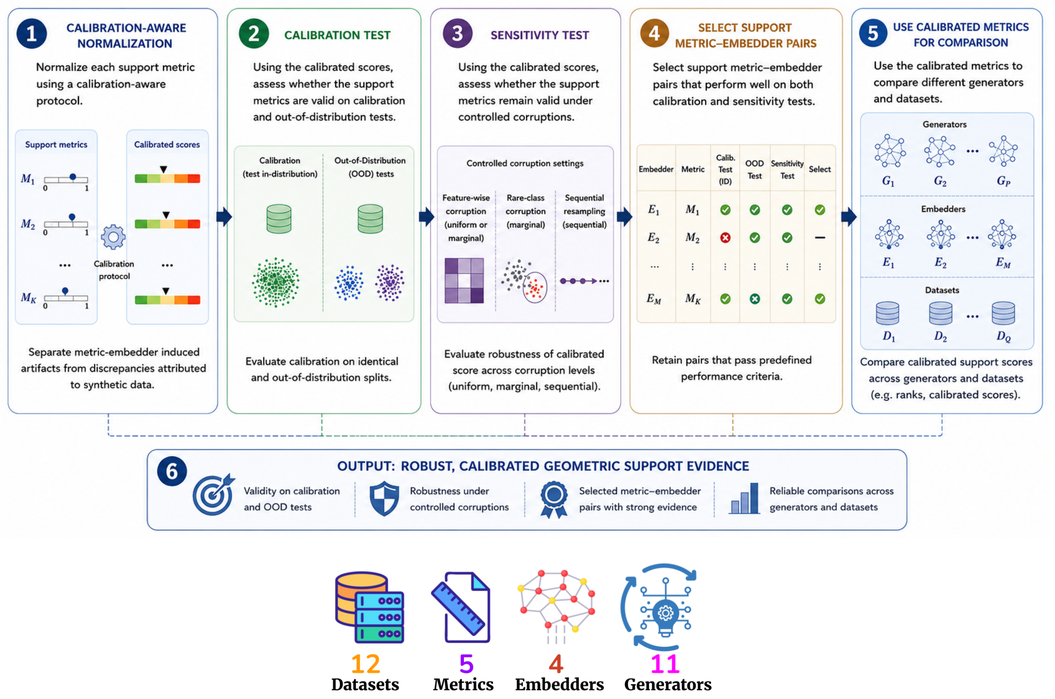
Contribution 1: *Baseline* Calibration Protocol



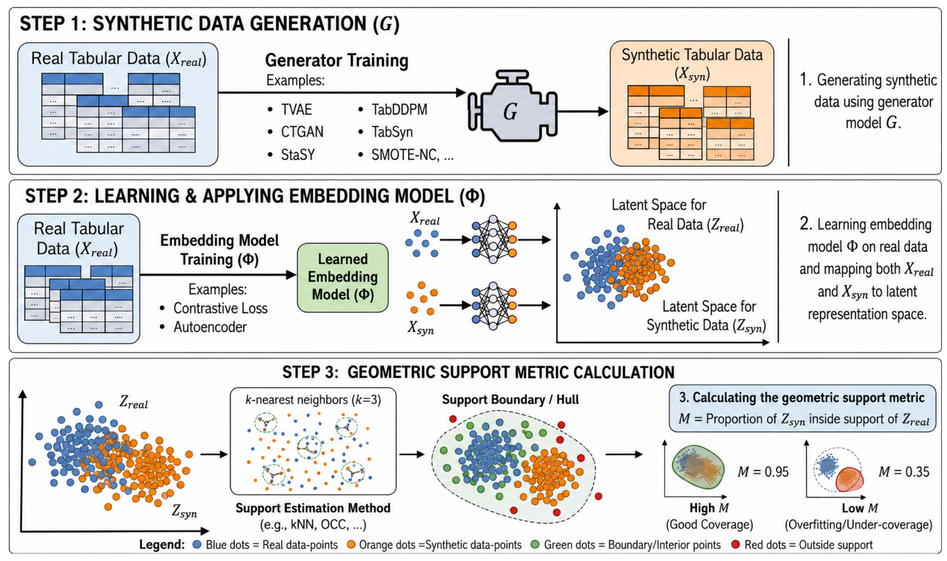
Contribution 2: Sensitivity Validation Tests



Calibration-Aware Geometric Support Framework

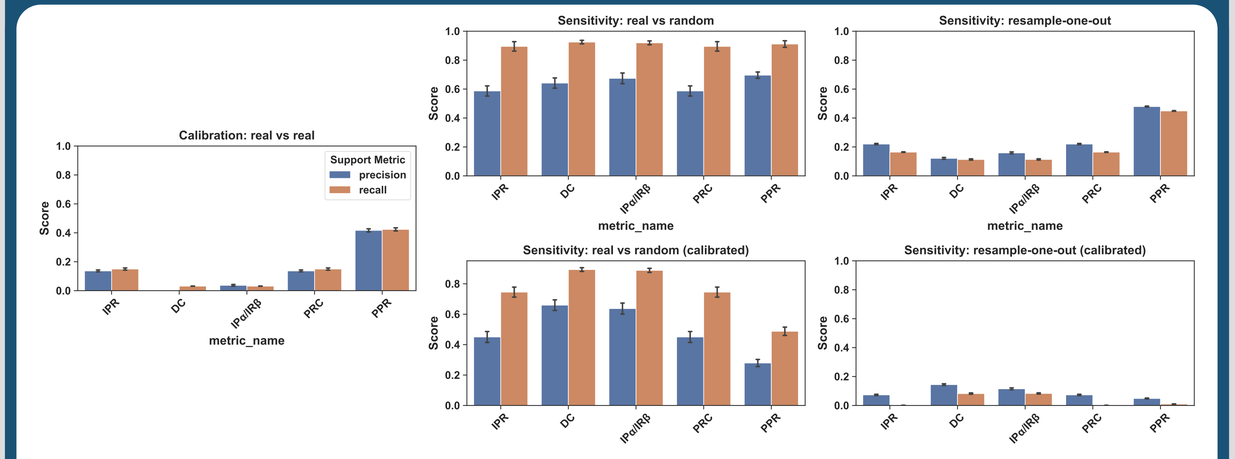


Current Geometric Support Framework



Geometric Support Metric ($\mathcal{M}$) workflow for synthetic tabular data. The process involves generating synthetic data ($Xsyn$) from real data ($Xreal$) using generator G (Step 1), learning an embedding model Φ to map both datasets into latent representations $Zreal$ and $Zsyn$ (Step 2), and finally calculating ($\mathcal{M}$) as the proportion of $Zsyn$ points falling within the estimated geometric support boundary (hull) of $Zreal$ (Step 3).

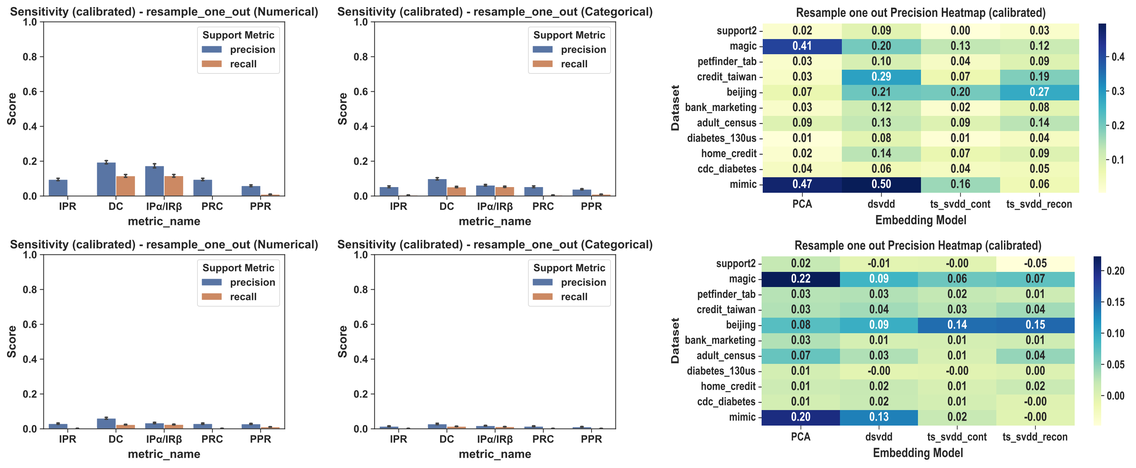
Baseline Bias: Calibration-Aware Scores



Calibration and baseline sensitivity tests: comparison of uncalibrated and calibrated aggregated support scores. Left: baseline calibration, illustrating the bias of each metric. Middle and right: aggregated support scores for uncalibrated (top) and calibrated (bottom) sensitivity tests: real data versus uniform resamples (middle) and one-out resampling perturbations (right). Error bars indicate the standard deviation across datasets and embedding models.

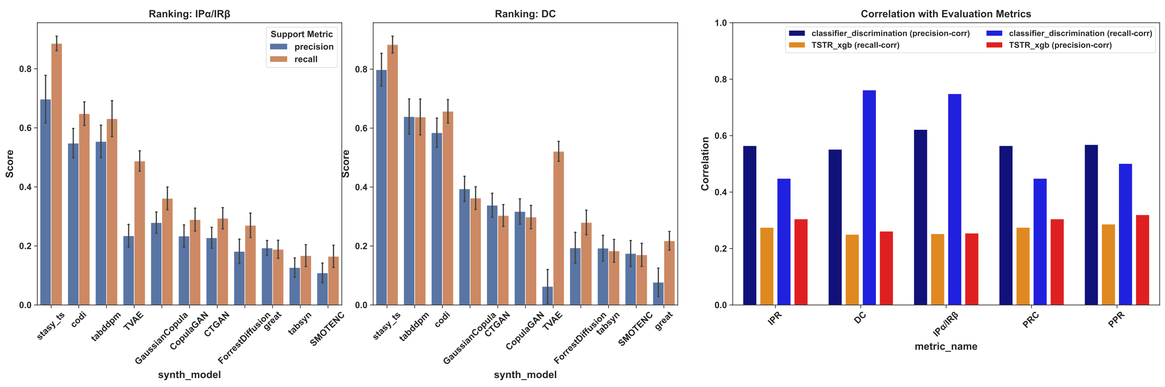
Raw geometric support scores can be dominated by embedder-induced bias
Calibration materially improves interpretability and ranking stability

Sensitivity Validation Tests: Tabular-Specific Shifts



Left & middle: One-out sensitivity tests using resampling (numerical & categorical). Calibrated support scores in the top row are based on uniform-resampling, while those in the bottom row use marginal-resampling. Right: support heatmaps showing the interaction between datasets and embedding models.

Calibration-Aware Generator Support Benchmark



Left & middle: the aggregated calibrated support scores of DC, and IP_α/IR_β on different tabular synthesizers. The standard deviation bars show variability of datasets and embedding models. Right: the correlation between calibrated geometric support and (1) classifier-discriminate and (2) machine-learning utility evaluations. Strong correlation with classifier discrimination confirms diagnostic utility; weak MLU correlation validates sample-level granularity.

Takeaways

- Support metrics are **not** black-box tools; they are pipeline-dependent and require dataset-specific calibration.
- Our protocol separates **baseline representation bias** from actual **synthetic data discrepancies**.
- Benchmarking reveals current metrics favor majority modes and **struggle with tabular conditional dependencies**.