mostlyai-qa: Quality Assurance for Synthetic Data

Michael Platzer MICHAEL.PLATZER@MOSTLY.AI

Mario Scriminaci MARIO.SCRIMINACI@MOSTLY.AI

Paul Tiwald PAUL.TIWALD@MOSTLY.AI

Abstract

We introduce `mostlyai-qa`, an open-source Python package for assessing synthetic data. The unique contribution of this toolkit lies in its ease of use, its broad support for mixed-type tabular data, and its combined assessment of fidelity and novelty. Using holdout data as a *north star* for quality, it produces a detailed, self-contained HTML report with a rich set of metrics and visualizations, allowing easy examination of synthetic data quality. The toolkit handles mixed-type data, such as numerical, categorical, datetime, and text data, along with multi-sequence time-series data as well as any contextual data, if available. The assessment is performed purely based on the provided data samples; therefore, no knowledge or assumptions about the underlying data generation mechanism are required. This facilitates benchmarking of emerging data synthesizers in the field. The package is available at `https://github.com/mostly-ai/mostlyai-qa/`, released under the Apache License v2, and will be continuously supported.

Keywords: synthetic data, generative models, data quality assessment, privacy-preserving data, mixed-type data analysis, privacy-utility trade-off

1 Introduction

Generative AI is rapidly gaining prominence, not only within unstructured data domains such as images and natural language but also within structured and semi-structured domains, that are common for proprietary data assets within organizations. The ability to produce an unlimited number of novel samples, that generalize beyond previously observed instances, and that help downstream learning, for both humans and machines alike, is of significant value. Such capability allows, among others, the privacy-safe sharing of micro-data across organizations, the sampling of otherwise underrepresented groups, the simulation of rare scenarios, and the filling of data gaps (Assefa et al., 2020; Jordon et al., 2022; van Breugel et al., 2024). Fulfilling this promise largely depends upon the quality of the generated synthetic data. The key question is whether synthetic samples are indeed truly novel as well as faithfully representative of the original statistics.

Although numerous evaluation frameworks have been introduced before (Howe et al., 2017; Lu et al., 2019; Platzer and Reutterer, 2021; Chundawat et al., 2022b,a; Alaa et al., 2022; Espinosa and Figueira, 2023; Task et al., 2023; Hudovernik et al., 2024), each presenting a different set of metrics, there remains a scarcity of well-maintained open-source packages that quantify and visualize synthetic data quality. See Table 1 for a high-level tool comparison. In particular, there is a lack of software packages that address both the fidelity and novelty of samples at the same time. Note that it is easy to excel in one dimension while neglecting the other. For instance, merely copying original samples yields high accuracy without being novel, while generating entirely random samples scores high on novelty without being accurate. The true challenge of privacy-safe synthetic data lies in the generation of data that is both accurate *and* novel. Thus, any quality assurance for synthetic data has to measure both of these dimensions.

Further, we aim to meet the rising demand for evaluating mixed-type synthetic data, characterized by: 1) diverse feature types, including numerical, categorical, datetime, and text; 2) missing values; 3) variable row counts per sample, accommodating multi-sequence, multivariate time-series data[1]; and 4) the incorporation of contextual data.

2 Concepts

Platzer and Reutterer (2021) introduced a holdout-based empirical assessment framework for mixed-type synthetic data, arguing that synthetic samples should emulate holdout samples — original samples excluded from the synthesis process. These holdout samples serve as a *north star* for privacy-preserving data synthesis, expecting models to produce novel samples that reflect the underlying data distribution without direct replication. Accordingly, synthetic samples should be as close to training samples as holdout samples are, but not closer. This approach, akin to the use of holdout samples for supervised learning, enables the evaluation of a generative model's ability to generalize underlying patterns rather than merely memorizing specific training samples.

Our toolkit builds upon that framework and introduces three sets of metrics - Accuracy, Similarity, and Distances - which will be described in the following. Accuracy quantifies lower-dimensional, and similarity higher-dimensional fidelity; the set of distance metrics helps to gauge the novelty of samples.

2.1 Accuracy

The accuracy metrics assess how closely the lower-dimensional marginal distributions of synthetic data align with those of the original, with 100% indicating a perfect match. This comparison is made for univariate and bivariate distributions across all attributes (see Figures 3 and 4) and averaged to produce a single overall metric. Following Platzer and Reutterer (2021), we discretize each attribute into a fixed number of 10 categories to facilitate comparison across mixed types. Numerical and datetime attributes are binned by deciles of the original data, ensuring roughly equal-sized groups. For categorical attributes, we retain the top categories by frequency, and for text, we tokenize and focus on the most common tokens.

This approach offers consistency across attribute types. Additionally, the overall accuracy metric is decomposable into 1-way and 2-way frequency tables, which are visualized, making it easily interpretable also for non-statisticians. The greater the discrepancies between the plotted distributions, the lower the accuracy score. To achieve a high overall accuracy, each contributing distribution must align closely with the original. However, due to sampling noise with finite samples, some discrepancies are inevitable. By calculating the expected accuracy for a theoretical holdout dataset based on the original distributions and sample size, we provide a reference benchmark. Rather than aiming for 100% accuracy, the

[1] Multi-sequence time-series data is the predominant structure for behavioral data, where multiple events for multiple individuals are recorded.

goal is for synthetic samples to match this benchmark closely, indicating similarity to the training samples akin to holdout samples.

When contextual data is present, the toolkit will report the accuracy of bivariate distributions between contextual and target attributes, enabling assessment of whether these relationships are well-preserved in the synthetic data.

For sequential data, the toolkit will evaluate attribute coherence between two randomly selected successive records within each sample. This allows assessment of whether the original sample sequences' autocorrelations are faithfully reproduced in the synthetic data.

2.2 Similarity

Complementing accuracy, we report another set of metrics that assess the similarity of distributions. Rather than analyzing the easy-to-interpret lower-dimensional marginals, the focus shifts to the high-dimensional full joint distributions. Direct analysis of high-dimensional distributions is not feasible due to the curse of dimensionality, so we use an alternative approach: each tabular sample is converted into a string of values, that is then mapped into an informative embedding space using a pre-trained language model. While the choice of language model is flexible, we specifically opted for `all-MiniLM-L6-v2`[2] as it is a lightweight, compute-efficient universal model. In this embedding space, we compare the centroids of the synthetic and training samples using cosine similarity, aiming for a high similarity score (with an upper bound of 1). However, to account also here for sampling variance, we use the cosine similarity between the training and holdout centroids as a *north star* reference, ensuring that synthetic samples are close to the training distribution without exceeding the similarity expected due to natural sampling noise.

To enhance interpretability, we provide a visualization of the embedded samples and their centroids, projected into a lower-dimensional space using Principal Component Analysis (PCA) (see Figure 6).

In addition to cosine similarity, we leverage the embedding space to train a discriminative model that indicates whether synthetic samples are truly indistinguishable from training samples. If certain properties of the synthetic samples (e.g., implausible attribute combinations) reveal them as synthetic rather than real, the area-under-the-curve (AUC) metric quantifies this distinguishability.

2.3 Distances

Synthetic samples shall resemble *novel* samples from the original distribution rather than simply replicating seen samples. Consequently, they are expected to be just as close to training samples as to holdout samples.

Thus, we assess the novelty of the synthetic data by examining distances between samples within the high-dimensional embedding space introduced in Section 2.2. For each synthetic sample, we calculate the distance to its closest record (DCR) among the training samples. This nearest-neighbor distance is expected to vary depending on whether the sample is a synthetic inlier or outlier. Therefore, absolute distances alone cannot reliably indicate novelty; instead, we need to contextualize these values by comparing them

[2] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2/`

to the same distances calculated for an equally sized holdout dataset. This comparison is performed for both the average DCR, which we report as a metric, and the overall DCR distribution, which is visualized (see Figure 7).

The need to compare to the corresponding holdout metrics also applies when checking for identical matches—cases where synthetic and original records are identical across all attributes. It is crucial to note that the existence of identical matches does not inherently indicate a lack of novelty. If the original data contains duplicates, we expect and require a similar proportion of matches in the synthetic dataset. Attempting to enforce novelty by removing individual records is not sufficient and may inadvertently risk exposing original records (Hann, 2024).

3 Empirical Demonstration

By splitting the original data into training and holdout samples and, subsequently, generating multiple synthetic datasets based on the training data, we can effectively compare quality across various methods. The chart below visualizes key metrics relative to their holdout-based reference metrics for the UCI Adult Census dataset (Dua and Graff, 2019), as synthesized and published in Platzer and Reutterer (2021). The closer a synthesizer approaches the *north star* reference point at `(1, 1)`, the better its privacy-utility trade-off. As illustrated, this trade-off applies to AI-based data synthesizers just as it does to traditional perturbation techniques. These metrics enable effective comparisons both within and across groups of techniques.

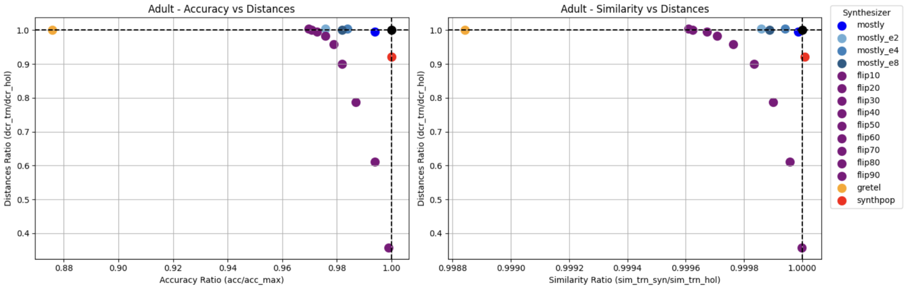


Figure 1: Comparison of synthesizers for UCI Adult Census dataset.

4 Conclusions

It is the need to measure fidelity and novelty, to accommodate for heterogeneity in data structure, and the objective to effectively inform a broad audience that drives our motivation to introduce this new Python toolkit. We hope to contribute to the continued standardization around the quality assessment of synthetic data.

Acknowledgements

We wish to convey our sincere appreciation to Tobias Hann, Radu Rogojanu, Lukasz Kolodziejczyk, Michael Druk, Shuang Wu, Alex Ichim, Andrey Sidorenko, Ivona Krchova, among others, for their essential contributions to the development of the `mostlyai-qa` toolkit. Furthermore, our gratitude extends to the community of users whose ongoing feedback and support have been paramount in the refinement and advancement of this toolkit. Their insights have been instrumental in tailoring `mostlyai-qa` to better address the requirements of practitioners engaged in synthetic data quality assessment. We also recognize the developers and maintainers of the open-source libraries upon which `mostlyai-qa` relies. Specifically, we appreciate the efforts of the teams responsible for `plotly`, `scikit-learn` and `transformers`.

Appendix A. Metrics Overview

- **Accuracy**: Accuracy is defined as (100% - Total Variation Distance), for each distribution, and then averaged across.
  - `overall`: Overall accuracy of synthetic data, i.e. average across univariate, bivariate and coherence.
  - `univariate`: Average accuracy of discretized univariate distributions.
  - `bivariate`: Average accuracy of discretized bivariate distributions.
  - `coherence`: Average accuracy of discretized coherence distributions. Only applicable for sequential data.
  - `overall_max`: Expected overall accuracy of a same-sized holdout. Serves as reference for `overall`.
  - `univariate_max`: Expected univariate accuracy of a same-sized holdout. Serves as reference for `univariate`.
  - `bivariate_max`: Expected bivariate accuracy of a same-sized holdout. Serves as reference for `bivariate`.
  - `coherence_max`: Expected coherence accuracy of a same-sized holdout. Serves as reference for `coherence`.
- **Similarity**: All similarity metrics are calculated within an embedding space.
  - `cosine_similarity_training_synthetic`: Cosine similarity between training and synthetic centroids.
  - `cosine_similarity_training_holdout`: Cosine similarity between training and holdout centroids. Serves as reference for `cosine_similarity_training_synthetic`.
  - `discriminator_auc_training_synthetic`: Cross-validated AUC of a discriminative model to distinguish between training and synthetic samples.
  - `discriminator_auc_training_holdout`: Cross-validated AUC of a discriminative model to distinguish between training and holdout samples. Serves as reference for `discriminator_auc_training_synthetic`.
- **Distances**: All distance metrics are calculated within an embedding space. An equal number of training and holdout samples is considered.
  - `ims_training`: Share of synthetic samples that are identical to a training sample.
  - `ims_holdout`: Share of synthetic samples that are identical to a holdout sample. Serves as reference for `ims_training`.
  - `dcr_training`: Average L2 nearest-neighbor distance between synthetic and training samples.
  - `dcr_holdout`: Average L2 nearest-neighbor distance between synthetic and holdout samples. Serves as reference for `dcr_training`.
  - `dcr_share`: The share of synthetic samples that are closer to a training sample than to a holdout sample. This shall not be significantly larger than 50%.

Appendix B. Tool Comparison

Python package	License	HTML	Plots	Metrics	Novelty	Data
`mostlyai-qa` (2024)	Apache	✓	✓	✓	✓	flexible
`ydata-profiling` (2024)	MIT	✓	✓	-	-	flexible
`sdmetrics` (2024)	MIT	-	✓	✓	✓	flexible
`synthcity` (2023)	Apache	-	✓	✓	✓	flexible
`sdnist` (2023)	permissive	✓	✓	✓	∼	fixed

Table 1: Comparison across open-source Python libraries for assessing synthetic data.

Appendix C. Basic Usage

The presented toolkit for evaluating the quality of synthetic data requires Python version 3.10 or later, and can be easily installed using `pip`:

```
pip install -U mostlyai-qa
```

Once installed, its main interface is the ‘report‘, which expects the data samples to be provided as `pandas` DataFrames:

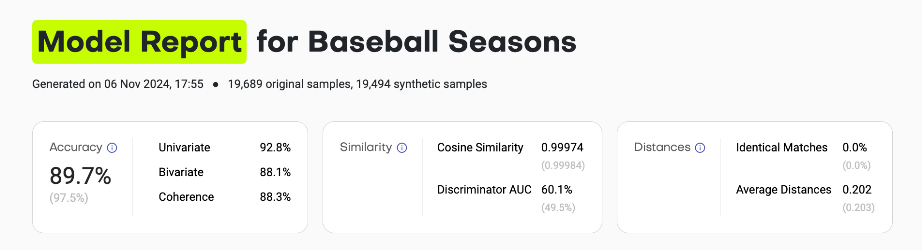
```
from mostlyai import qa

# analyze single-table data
report_path, metrics = qa.report(
    syn_tgt_data=synthetic_df,
    trn_tgt_data=training_df,
    hol_tgt_data=holdout_df,
)

# analyze sequential data with context
report_path, metrics = qa.report(
    syn_tgt_data=synthetic_df,
    trn_tgt_data=training_df,
    hol_tgt_data=holdout_df,
    syn_ctx_data=synthetic_context_df,
    trn_ctx_data=training_context_df,
    hol_ctx_data=holdout_context_df,
    ctx_primary_key="id",
    tgt_context_key="user_id",
)
```

Additional usage examples, along with their corresponding HTML reports, are available in the GitHub repository.

Appendix D. HTML Report Example



Model Report for Baseball Seasons

Generated on 06 Nov 2024, 17:55 • 19,689 original samples, 19,494 synthetic samples

Accuracy ⓘ
89.7%
(97.5%)

Univariate	92.8%
Bivariate	88.1%
Coherence	88.3%

Similarity ⓘ

Cosine Similarity	0.99974 (0.99984)
Discriminator AUC	60.1% (49.5%)

Distances ⓘ

Identical Matches	0.0% (0.0%)
Average Distances	0.202 (0.203)

Figure 2: Metrics summary.

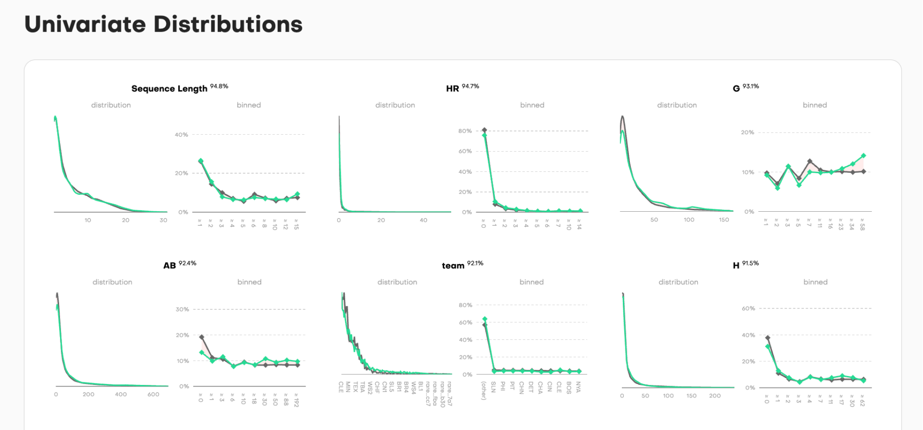


Figure 3: Univariate distributions and their accuracies.

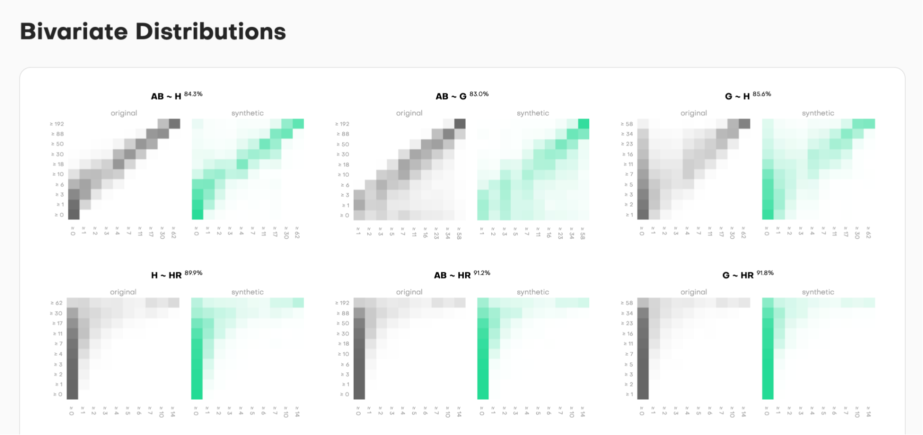
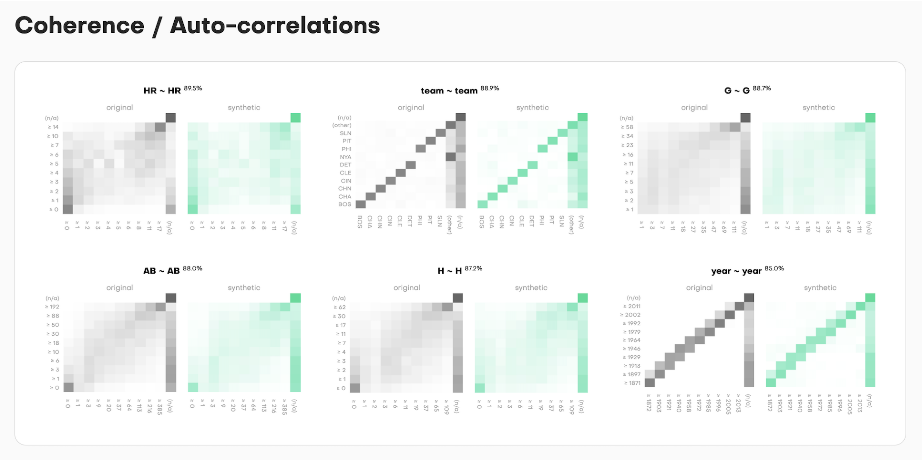
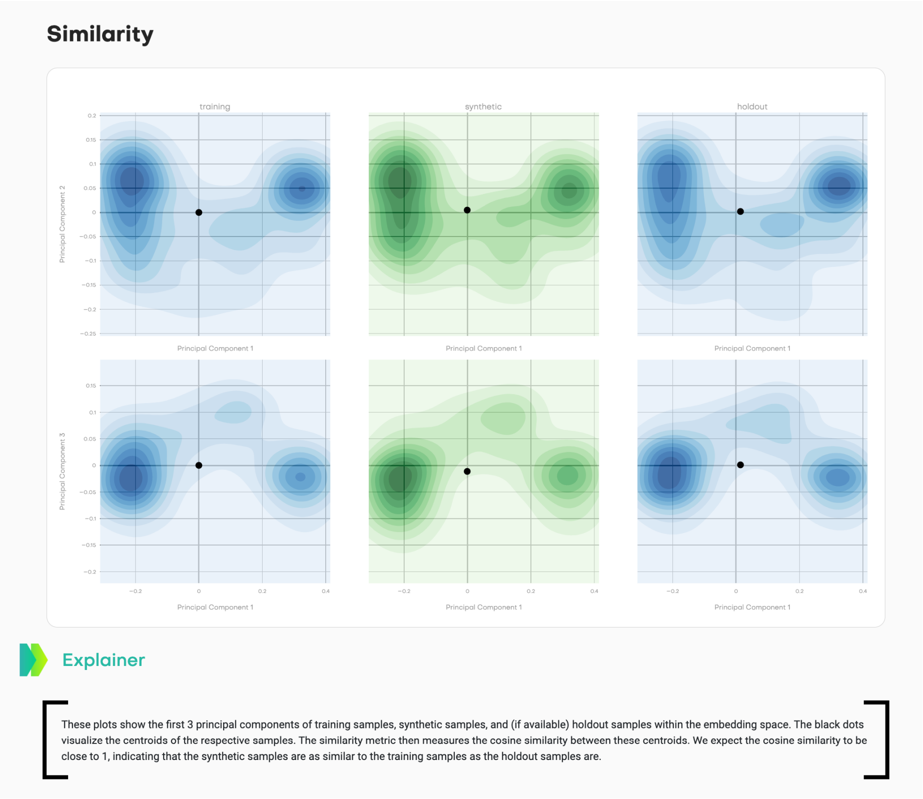


Figure 4: Bivariate distributions and their accuracies.



Coherence / Auto-correlations

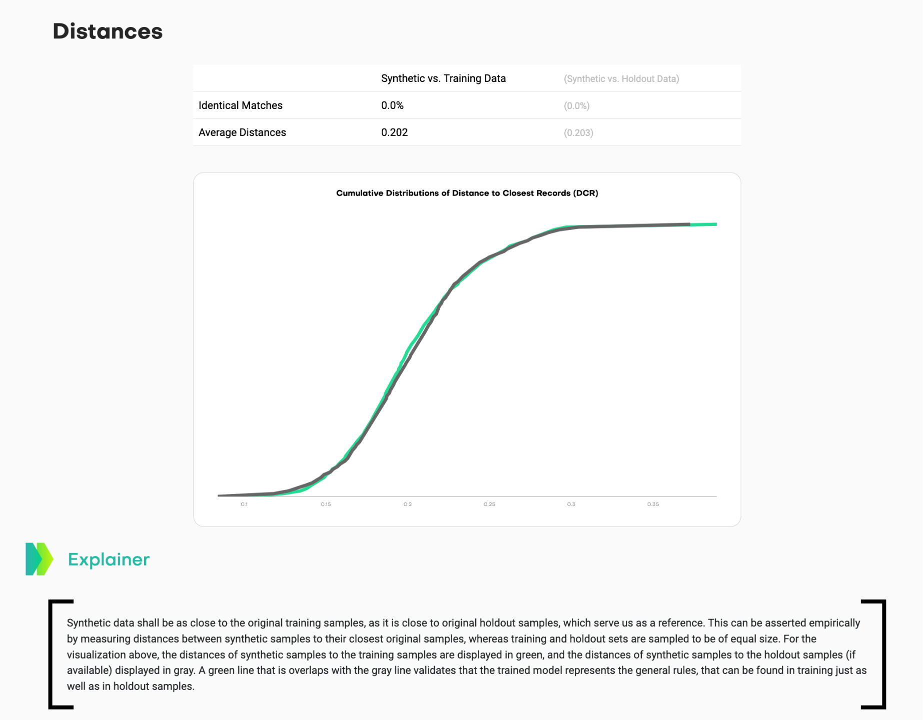
Figure 5: Coherence distributions and their accuracies.



Similarity

These plots show the first 3 principal components of training samples, synthetic samples, and (if available) holdout samples within the embedding space. The black dots visualize the centroids of the respective samples. The similarity metric then measures the cosine similarity between these centroids. We expect the cosine similarity to be close to 1, indicating that the synthetic samples are as similar to the training samples as the holdout samples are.

Figure 6: Similarity within PCA-projected embedding space.



Distances

	Synthetic vs. Training Data	(Synthetic vs. Holdout Data)
Identical Matches	0.0%	(0.0%)
Average Distances	0.202	(0.203)

Cumulative Distributions of Distance to Closest Records (DCR)
0.1
0.15
0.2
0.25
0.3
0.35
Explainer

Synthetic data shall be as close to the original training samples, as it is close to original holdout samples, which serve us as a reference. This can be asserted empirically by measuring distances between synthetic samples to their closest original samples, whereas training and holdout sets are sampled to be of equal size. For the visualization above, the distances of synthetic samples to the training samples are displayed in green, and the distances of synthetic samples to the holdout samples (if available) displayed in gray. A green line that is overlaps with the gray line validates that the trained model represents the general rules, that can be found in training just as well as in holdout samples.

Figure 7: Distribuions of distance to closest records (DCRs) for assessing novelty.

References

A. Alaa, B. Van Breugel, E. S. Saveliev, and M. van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8, 2020.

V. S. Chundawat, A. K. Tarun, M. Mandal, M. Lahoti, and P. Narang. Tabsyndex: A universal metric for robust evaluation of synthetic tabular data. *arXiv preprint arXiv:2207.05295*, 2022a.

V. S. Chundawat, A. K. Tarun, M. Mandal, M. Lahoti, and P. Narang. A universal metric for robust evaluation of synthetic tabular data. *IEEE Transactions on Artificial Intelligence*, 5(1):300–309, 2022b.

DataCebo. SDMetrics. `https://github.com/sdv-dev/SDMetrics`, 2024.

D. Dua and C. Graff. UCI machine learning repository: Adult data set, 2019. URL `https://archive.ics.uci.edu/ml/datasets/adult`. University of California, Irvine, School of Information and Computer Sciences.

E. Espinosa and A. Figueira. On the quality of synthetic generated tabular data. *Mathematics*, 11(15):3278, 2023.

T. Hann. Why removing identical matches in synthetic data risks privacy: The swiss cheese problem. https://mostly.ai/blog/why-removing-identical-matches-in-synthetic-data-risks-privacy-the-swiss-cheese-problem, April 2024. Blog post.

B. Howe, J. Stoyanovich, H. Ping, B. Herman, and M. Gee. Synthetic data for social good. *arXiv preprint arXiv:1710.08874*, 2017.

V. Hudovernik, M. Jurkovič, and E. Štrumbelj. Benchmarking the fidelity and utility of synthetic relational data. *arXiv preprint arXiv:2410.03411*, 2024.

J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, and A. Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

P.-H. Lu, P.-C. Wang, and C.-M. Yu. Empirical evaluation on synthetic data generation with generative adversarial network. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, pages 1–6, 2019.

MOSTLY AI. mostlyai-qa. https://github.com/mostly-ai/mostlyai-qa, 2024.

M. Platzer and T. Reutterer. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data*, 4:679939, 2021.

Z. Qian, B.-C. Cebere, and M. van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023. URL https://arxiv.org/abs/2301.07573.

C. Task, K. Bhagat, and G. Howarth. SDNist. https://github.com/usnistgov/SDNist, 2023. URL https://doi.org/10.18434/mds2-2943.

B. van Breugel, T. Liu, D. Oglic, and M. van der Schaar. Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*, pages 1–14, 2024.

YData. ydata-profiling. https://github.com/ydataai/ydata-profiling, 2024.