ANITA

Anonymous big data A project funded by FFG

Simulation Study Results

Deliverable D4.2

Author(s): Michael Platzer, Klaudius Kalcher

Reviewer(s): Stefan Vamosi

Document version: 0.2
Date: 07.06.2021

Federal Ministry Republic of Austria Climate Action, Environment, Energy, Mobility, Innovation and Technology

Programme **"ICT of the Future"** - the initiative of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK)

Disclaimer

This deliverable describes the work and findings of the AI-Based Privacy-Preserving Big Data Sharing for Market Research (Anonymous Big Data (ANITA)) project.

The authors of this document have made every effort to ensure that its content was accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this deliverable are responsible for any possible errors or omissions as well as for any results and actions that might occur as a result of using the content of this document.

Table of contents

SIMULATION STUDY RESULTS 1
DISCLAIMER 2
TABLE OF CONTENTS 3
1 SUMMARY 4
2 SDGYM BENCHMARKS 5
3 ASSESSMENT FRAMEWORK BENCHMARKS 6

1 Summary

The simulation study leveraged the Virtual Data Lab (see D4.1.) to benchmark the three included synthesizers across the four included mixed-type sequential datasets (CDNOW, BERKA, MLB, RETAIL) across all introduced accuracy and privacy metrics. In addition, MOSTLY AI's proprietary synthetic data solution has been integrated into these benchmarks via the provided virtual data lab interface. All computations were performed on Google cloud GPU resources.

index	synthesizer	TVD univariate	L1D univariate	L1D bivariate	L1D 3-way	L1D 4-way	L1D Users per Category	L1D Categories per User	DCR test	NNDR test
berka	IdentitySynthesizer	0.01153	0.02964	0.05447	0.06299	0.09058	0.02305	0.0176	FAILED	FAILED
berka	ShuffleSynthesizer	0.00785	0.0226	0.26474	0.44412	0.56838	0.74125	0.7281	PASSED	PASSED
berka	FlatAutoEncoderSynthesizer	0.12037	0.34743	0.54751	0.69484	0.75032	1.61395	0.94685	PASSED	PASSED
berka	MOSTLY	0.02564	0.09338	0.17975	0.27691	0.32319	0.2188	0.1553	PASSED	PASSED
cdnow	IdentitySynthesizer	0.01749	0.04909	0.08113	0.10386	0.17095	0.04221	0.03533	FAILED	FAILED
cdnow	ShuffleSynthesizer	0.01478	0.04696	0.22987	0.4613	0.46676	0.27686	0.32323	PASSED	PASSED
cdnow	FlatAutoEncoderSynthesizer	0.31145	0.9025	1.2048	1.32703	1.38919	1.81962	0.49682	PASSED	PASSED
cdnow	MOSTLY	0.02093	0.07834	0.15366	0.25723	0.24986	0.2213	0.11664	PASSED	PASSED
mlb	IdentitySynthesizer	0.01165	0.0363	0.07706	0.1331	0.17534	0.0619	0.0259	FAILED	FAILED
mlb	ShuffleSynthesizer	0.0108	0.03795	0.27379	0.4406	0.6002	1.40715	1.05925	PASSED	PASSED
mlb	FlatAutoEncoderSynthesizer	0.3086	0.91435	1.26081	1.35781	1.37516	2.84045	1.33855	PASSED	PASSED
mlb	MOSTLY	0.02564	0.09338	0.17975	0.27691	0.32319	0.2188	0.1553	PASSED	PASSED

Key findings:

- IdentitySynthesizer and ShuffleSynthesizer exhibit best scores with respect to the univariate accuracy measures
- IdentitySynthesizer does not pass the privacy tests – this is as expected, and validates the proper functioning of the privacy tests
- ShuffleSynthesizer, which randomly shuffles all columns across all records, destroys the multi-variate information and thus results in worse scores for all except the univariate measures
- FlatAutoEncoderSynthesizer, which is a fully-connected Auto-Encoder adapted to sequential data, passes the privacy tests, however, achieves very poor accuracy results. It isn't able to capture the univariate statistics well, hence also yields poor scores for higher-level accuracy metrics, that are even lower than for the ShuffleSynthesizer. Note, that the FlatAutoEncoderSynthesizer was included into the Virtual Data Lab as a proof-of-concept, and as demonstration for the implementation of custom AI-based synthesizers.
- The MOSTLY synthesizer passes all privacy tests, and remains close to the higher-level statistical distributions. These include multi-variate relations, as well as the introduced coherence measures.

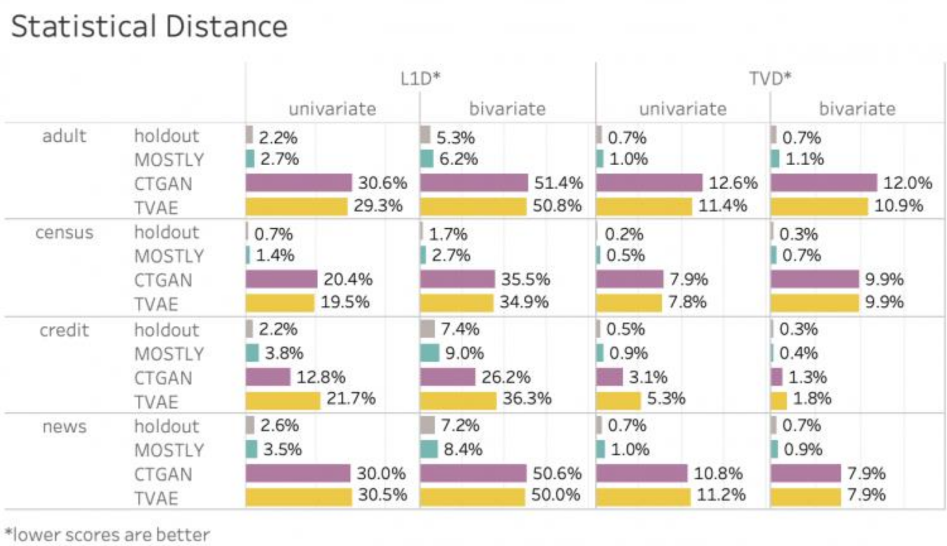
Note, that all further simulation results related to WP5 are presented together with the corresponding WP5 deliverables.

2 SDGym Benchmarks

In addition to the planned simulation study on top of Virtual Data Lab, an extensive benchmarking study was performed for non-sequential mixed-type datasets, on top of MIT's SDgym library. Detailed results were published at https://mostly.ai/2020/09/25/the-worlds-most-accurate-synthetic-data-platform/

- Six synthesizers:
  - CTGAN
  - MedGAN
  - TableGAN
  - TVAE
  - VEEGAN
  - MOSTLY

- Four single-table mixed-type datasets
  - **adult**: ~23'000 training records, 10'000 holdout records, with 14 mixed-type attributes and one binary target variable (24% class im**balan**ce)
  - **census**: ~200'000 training records, ~100'000 holdout records, with 40 mixed-type attributes and one binary target variable (6% class imbalance)
  - **credit**: ~265'000 training records, ~20'000 holdout records, with 29 numeric attributes and one binary target variable (0.17% class imbalance)
  - **news**: ~33'000 training records, 8'000 holdout records, with 58 mixed-type attributes and one numeric (log-transformed) target variable



ML Performance

Classification

		AUC	AUCPR	F1
adult	original	90.3%	76.1%	68.3%
	MOSTLY	89.7%	74.7%	66.5%
	CTGAN	84.9%	64.2%	54.8%
	TVAE	86.4%	67.0%	63.7%
census	original	92.4%	57.4%	49.5%
	MOSTLY	92.4%	56.3%	48.9%
	CTGAN	87.7%	42.0%	35.7%
	TVAE	89.9%	46.3%	42.3%
credit	original	90.5%	62.2%	50.7%
	MOSTLY	91.1%	58.6%	54.4%
	CTGAN	88.8%	51.2%	48.7%
	TVAE	72.1%	16.7%	5.0%

Higher scores are better.

Considered ML Models

- (All)
- AdaBoost
- DecisionTree
- LightGBM
- Linear
- LogReg
- MLP
- XGBoost

Regression

		R2	MAE*	RMSE*
news	original	0.133	0.623	0.836
	MOSTLY	0.124	0.633	0.840
	CTGAN	-0.004	0.688	0.899
	TVAE	-0.276	0.720	1.014

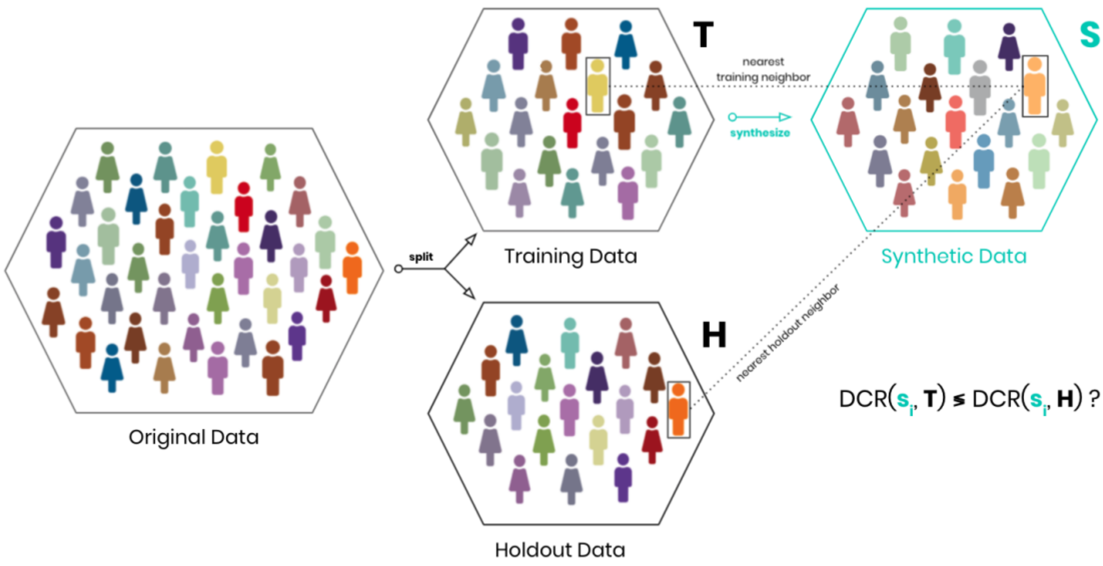
*Note: lower MAE and RMSE are better.

Key Findings:

- MOSTLY significantly outperforms existing open-source data synthesizers.
- This is true for the utility of downstream Machine Learning tasks, across a range of ML models and a range of ML accuracy metrics. But this is in particular true when it comes to the representativeness of the synthetic data measured as statistical distances.
- These findings are consistent across all benchmarked datasets.

3 Assessment Framework Benchmarks

We further developed an empirical holdout-based assessment framework for mixed-type synthetic data, and applied it to seven synthesizers, and four publicly available datasets. The key idea is to split an original dataset into a training dataset T, and a holdout dataset H, and derive the synthetic dataset S purely based on the training dataset T. This allows to then assess both the fidelity (i.e., the representativeness in terms of statistical distances) and the privacy of synthetic data in relation to a holdout data. In order to handle mixed-type data we proposed to discretize all variables and introduce an upper limit for the maximum cardinality.



The seven synthesizers included were

- CTGAN
- CopulaGAN
- GaussianCopula
- TVAE
- Gretel.ai
- MOSTLY
- Synthpop

The four mixed-type datasets were

- Adult: 48,842 rows, 15 attributes
- Credit-default: 30,000 rows, 24 attributes
- Marketing: 45,211 rows, 17 attributes
- Online-shoppers: 12,330 rows, 18 attributes

These are the key results of the study:

ANITA
Anonymous big data

[Fidelity] Average Total Variation Distance

		adult			bank-marketing			credit-default			online-shoppers		
		univariate (F1)	bivariate (F2)	three-way (F3)	univariate (F1)	bivariate (F2)	three-way (F3)	univariate (F1)	bivariate (F2)	three-way (F3)	univariate (F1)	bivariate (F2)	three-way (F3)
	Holdout	1.0%	1.6%	2.1%	1.0%	1.3%	1.7%	2.2%	2.2%	2.5%	2.2%	2.6%	2.7%
Synthesizers	CopulaGAN	13.1%	20.7%	26.4%	10.0%	13.8%	16.0%	16.4%	19.1%	21.4%	22.0%	29.4%	36.8%
	CTGAN	15.8%	20.9%	26.3%	10.6%	14.7%	17.2%	22.8%	25.0%	28.1%	24.5%	34.2%	43.2%
	GaussianCopula	28.9%	37.4%	45.0%	22.5%	29.5%	34.4%	30.2%	37.9%	43.9%	36.4%	52.5%	59.8%
	Gretel	4.2%	6.1%	8.1%	3.3%	5.4%	7.3%	11.5%	19.1%	25.1%	6.5%	9.8%	12.0%
	MOSTLY	1.3%	1.9%	2.4%	1.5%	2.0%	2.4%	3.8%	5.4%	5.8%	2.8%	3.2%	3.4%
	synthpop	0.6%	1.3%	1.9%	0.6%	1.1%	1.4%	1.3%	2.2%	2.8%	0.7%	1.3%	1.6%
	TVAE	27.7%	42.6%	49.3%	33.6%	46.6%	54.7%	47.0%	63.8%	73.0%	36.7%	50.9%	55.7%
Perturbate	Flip 10%	0.5%	1.7%	3.0%	0.6%	1.2%	1.7%	0.9%	4.0%	6.6%	0.6%	1.3%	1.9%
	Flip 20%	0.5%	2.8%	5.2%	0.5%	1.8%	2.9%	0.9%	7.3%	12.4%	0.6%	2.1%	3.4%
	Flip 30%	0.6%	3.9%	7.4%	0.5%	2.4%	3.9%	0.9%	10.1%	17.4%	0.6%	2.9%	4.7%
	Flip 40%	0.5%	4.7%	9.1%	0.5%	2.9%	4.8%	0.9%	12.7%	21.7%	0.5%	3.5%	5.8%
	Flip 50%	0.5%	5.4%	10.6%	0.5%	3.4%	5.7%	0.9%	14.8%	25.3%	0.5%	4.1%	6.8%
	Flip 60%	0.5%	6.1%	11.8%	0.5%	3.7%	6.3%	0.9%	16.6%	28.3%	0.6%	4.6%	7.6%
	Flip 70%	0.5%	6.6%	12.8%	0.5%	4.1%	6.8%	1.0%	18.0%	30.5%	0.5%	4.9%	8.2%
	Flip 80%	0.5%	6.9%	13.5%	0.5%	4.3%	7.1%	0.9%	19.0%	32.1%	0.6%	5.2%	8.6%
	Flip 90%	0.5%	7.1%	13.9%	0.5%	4.4%	7.3%	0.9%	19.6%	33.1%	0.6%	5.3%	8.8%

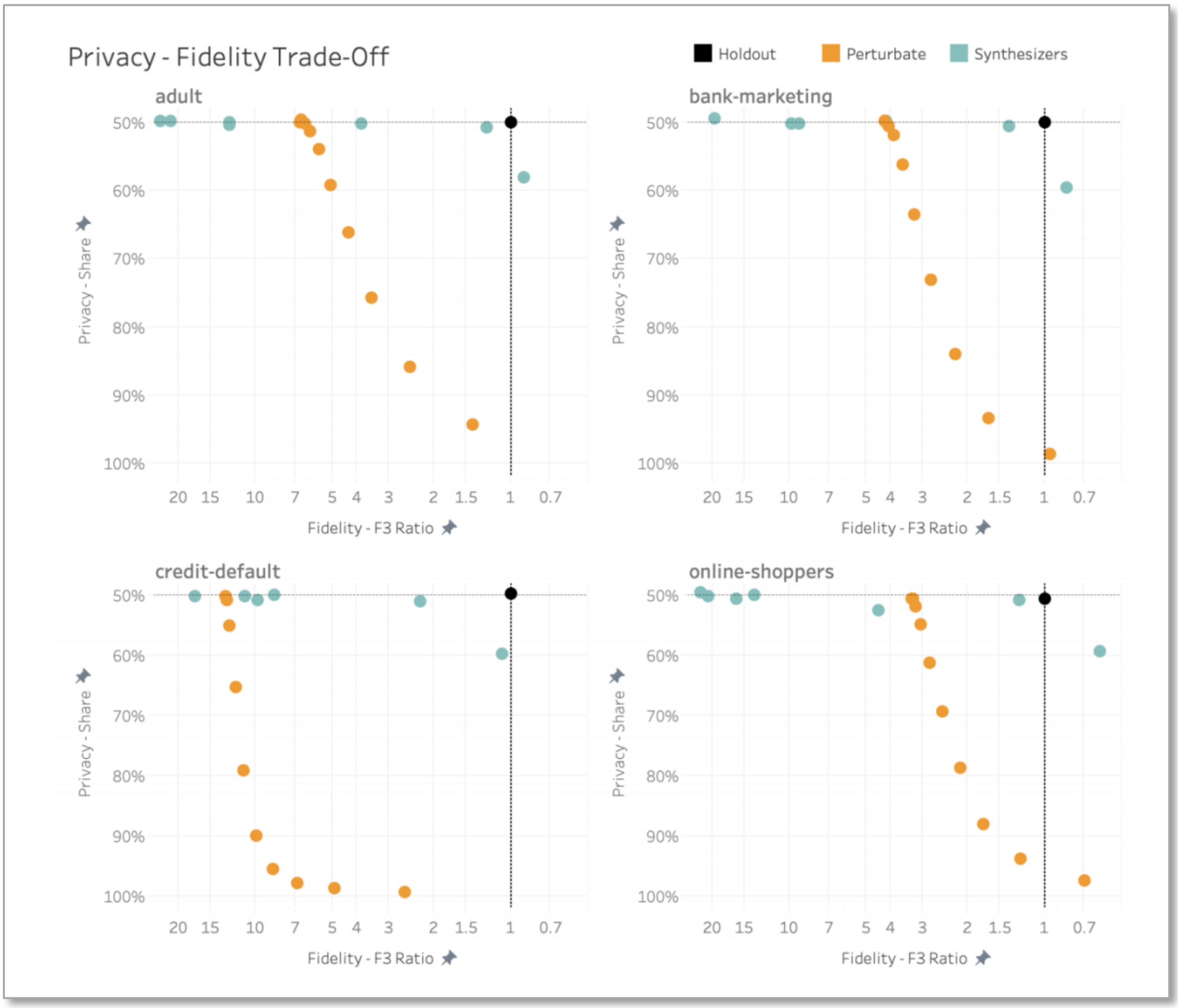
univariate c=100; bivariate c=10; three-way c=5

[Privacy] Distance to Closest Record - Training vs. Holdout

		adult			bank-marketing			credit-default			online-shoppers		
		Share	Avg DCR Train	Avg DCR Holdout	Share	Avg DCR Train	Avg DCR Holdout	Share	Avg DCR Train	Avg DCR Holdout	Share	Avg DCR Train	Avg DCR Holdout
	Holdout	50.0%	2.27	2.27	50.1%	3.57	3.58	49.8%	8.66	8.66	50.5%	4.28	4.29
Synthesizers	CopulaGAN	50.0%	4.19	4.19	50.2%	4.46	4.46	50.0%	12.04	12.04	49.8%	8.26	8.26
	CTGAN	50.4%	4.49	4.50	50.3%	4.61	4.61	50.1%	12.37	12.37	50.6%	8.59	8.60
	GaussianCopula	50.0%	5.54	5.54	49.6%	5.65	5.64	50.1%	13.82	13.82	49.5%	9.19	9.18
	Gretel	50.2%	2.49	2.49	49.9%	4.00	4.00	50.8%	10.95	10.97	52.4%	4.56	4.62
	MOSTLY	50.6%	2.34	2.35	50.7%	3.68	3.70	51.1%	9.81	9.83	50.9%	4.50	4.52
	synthpop	58.0%	2.14	2.33	59.6%	3.44	3.68	59.7%	8.97	9.26	59.3%	4.07	4.30
	TVAE	49.9%	3.89	3.89	51.3%	4.61	4.64	50.7%	14.31	14.32	50.2%	8.15	8.16
Perturbate	Flip 10%	94.3%	0.84	2.57	98.7%	0.96	3.76	99.4%	1.80	9.29	97.6%	0.92	4.32
	Flip 20%	85.8%	1.62	2.84	93.4%	1.89	3.92	98.8%	3.62	9.87	93.8%	1.83	4.43
	Flip 30%	75.8%	2.29	3.08	84.0%	2.71	4.06	98.0%	5.41	10.43	88.2%	2.73	4.64
	Flip 40%	66.2%	2.83	3.29	73.2%	3.38	4.18	95.7%	7.21	10.98	78.8%	3.40	4.60
	Flip 50%	59.2%	3.24	3.48	63.5%	3.87	4.27	90.1%	8.92	11.44	69.4%	3.97	4.67
	Flip 60%	54.0%	3.51	3.61	56.2%	4.16	4.34	79.2%	10.42	11.84	61.2%	4.39	4.74
	Flip 70%	51.4%	3.69	3.72	52.0%	4.34	4.39	65.3%	11.54	12.13	54.9%	4.63	4.76
	Flip 80%	50.3%	3.79	3.79	50.6%	4.41	4.43	55.0%	12.20	12.35	51.9%	4.76	4.81
	Flip 90%	49.8%	3.84	3.84	49.9%	4.45	4.45	50.8%	12.43	12.45	50.6%	4.83	4.84

c = 100

When visualized via a privacy-utility scatterplot, the clear relationship emerges between these two targets, whereas the holdout data serves as a north star, in terms of what is maximum achievable.



Further details are available at https://arxiv.org/abs/2104.00635 (preprint), or then in the upcoming paper by Platzer & Reutterer in [Frontiers in Big Data](#).

HOLDOUT-BASED FIDELITY AND PRIVACY ASSESSMENT OF MIXED-TYPE SYNTHETIC DATA

Michael Platzer
MOSTLY AI
Vienna, Austria
michael.platzer@mostly.ai

Thomas Reutterer
Vienna University of Economics and Business
Vienna, Austria
thomas.reutterer@wu.ac.at

ABSTRACT

AI-based data synthesis has seen rapid progress over the last several years, and is increasingly recognized for its promise to enable privacy-respecting high-fidelity data sharing. However, adequately evaluating the quality of generated synthetic datasets is still an open challenge. We introduce and demonstrate a holdout-based empirical assessment framework for quantifying the fidelity as well as the privacy risk of synthetic data solutions for mixed-type tabular data. Measuring fidelity is based on statistical distances of lower-dimensional marginal distributions, which provide a model-free and easy-to-communicate empirical metric for the representativeness of a synthetic dataset. Privacy risk is assessed by calculating the individual-level distances to closest record with respect to the training data. By showing that the synthetic samples are just as close to the training as to the holdout data, we yield strong evidence that the synthesizer indeed learned to generalize patterns and is independent of individual training records. We demonstrate the presented framework for seven distinct synthetic data solutions across four mixed-type datasets and compare these to more traditional statistical disclosure techniques. The results highlight the need to systematically assess the fidelity just as well as the privacy of these emerging class of synthetic data generators.