Generating Tabular Data using Diffusion Generative Models in a Hidden Space

Potemkina Alina

Supervisor: Mikhail Hushchyn Senior Research Fellow Ph.D.

HSE University
Faculty of Computer Science

Moscow, June 2025

Relevance

- Synthethic tabular data is extremely widespread and in demand in data science, and is widely used in a number of fields:
  - Extending training datasets to train machine learning models.
  - Protect private data by creating anonymized copies.
  - Filling in gaps in data to enhance its integrity and usefulness.
- The need to improve the quality of synthesized data to improve the performance of predictive models.
- The potential of using modern diffusion models to effectively solve the problems of modeling complex dependencies between features.

Problem statement

- The purpose of the work is to generate synthetic data that fully preserves the key statistical characteristics of the source data.
- This data should be anonymized in such a way as to exclude the possibility of restoring the original records or identifying the participants.

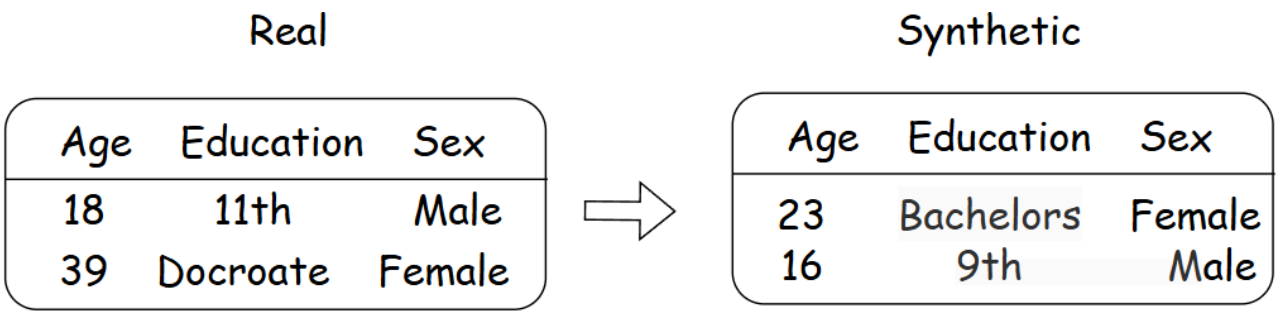


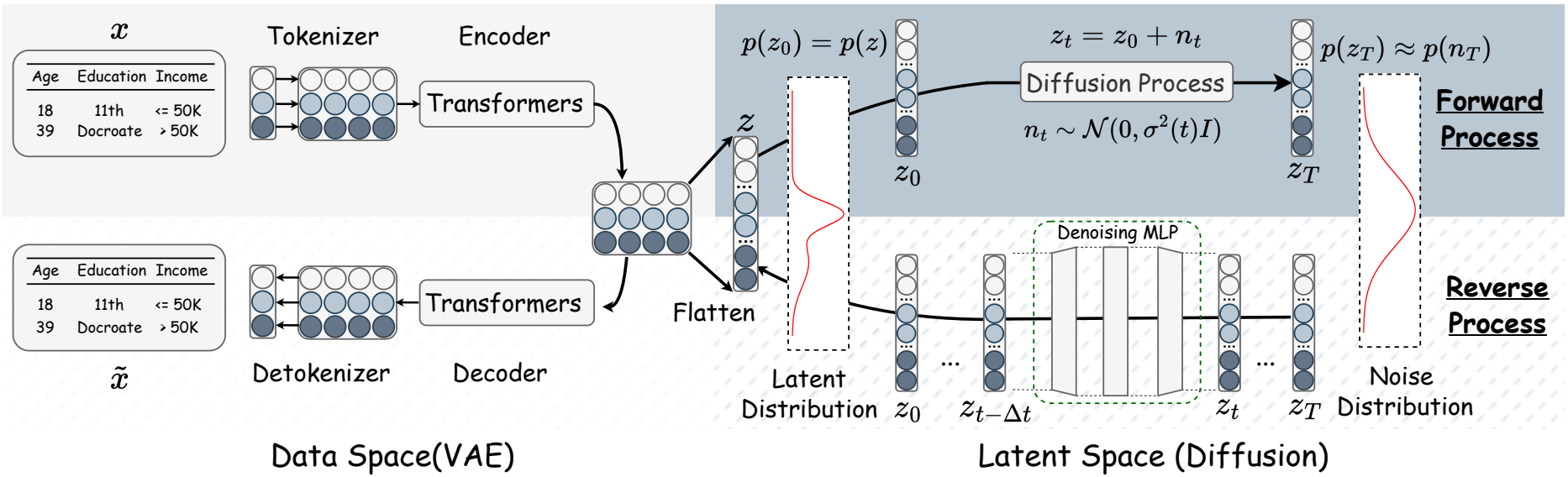
Figure: Example real and generated data

Previous research

- **CTGAN** and **TVAE** use different basic models (GAN and VAE), but the same normalization and conditional generation techniques to combat class unevenness and complex distributions of numerical features
- **GOGGLE** (Liu et al. 2023) implements graph neural networks to model dependencies between columns
- **GReaT** interprets a table row as a natural language sentence and trains a GPT model
- **STaSy** combines numerical and categorical features using diffusion processes
- **CoDi** and **TabDDPM** use separate diffusion processes for numerical and categorical features, but differ in the ways they combine models and the number of additional techniques.

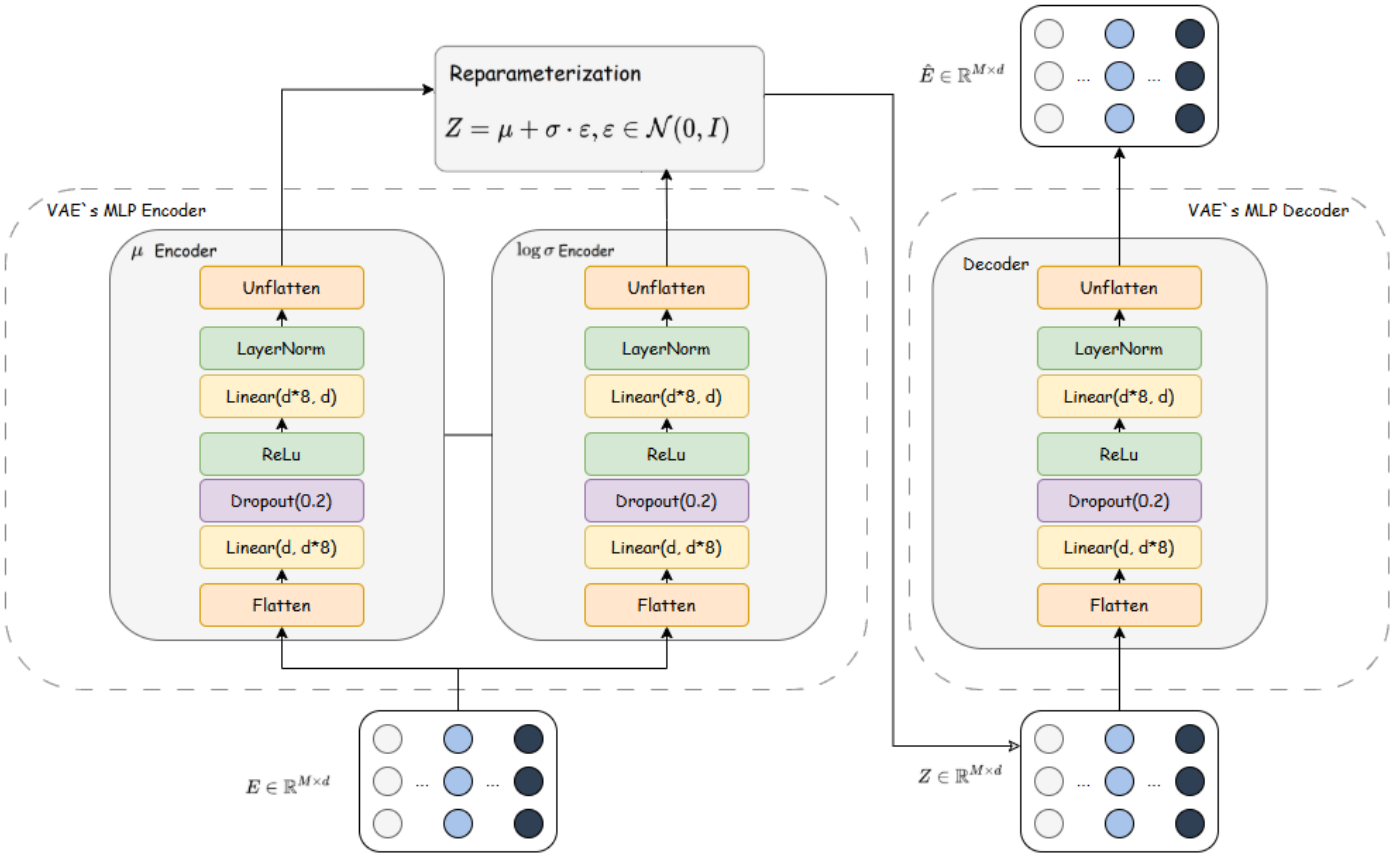
TabSyn Transformers-VAE

The TabSyn model uses transformer-based encoders and decoders to synthesize tabular data. It combines a diffusion model operating in a hidden space formed by VAE.



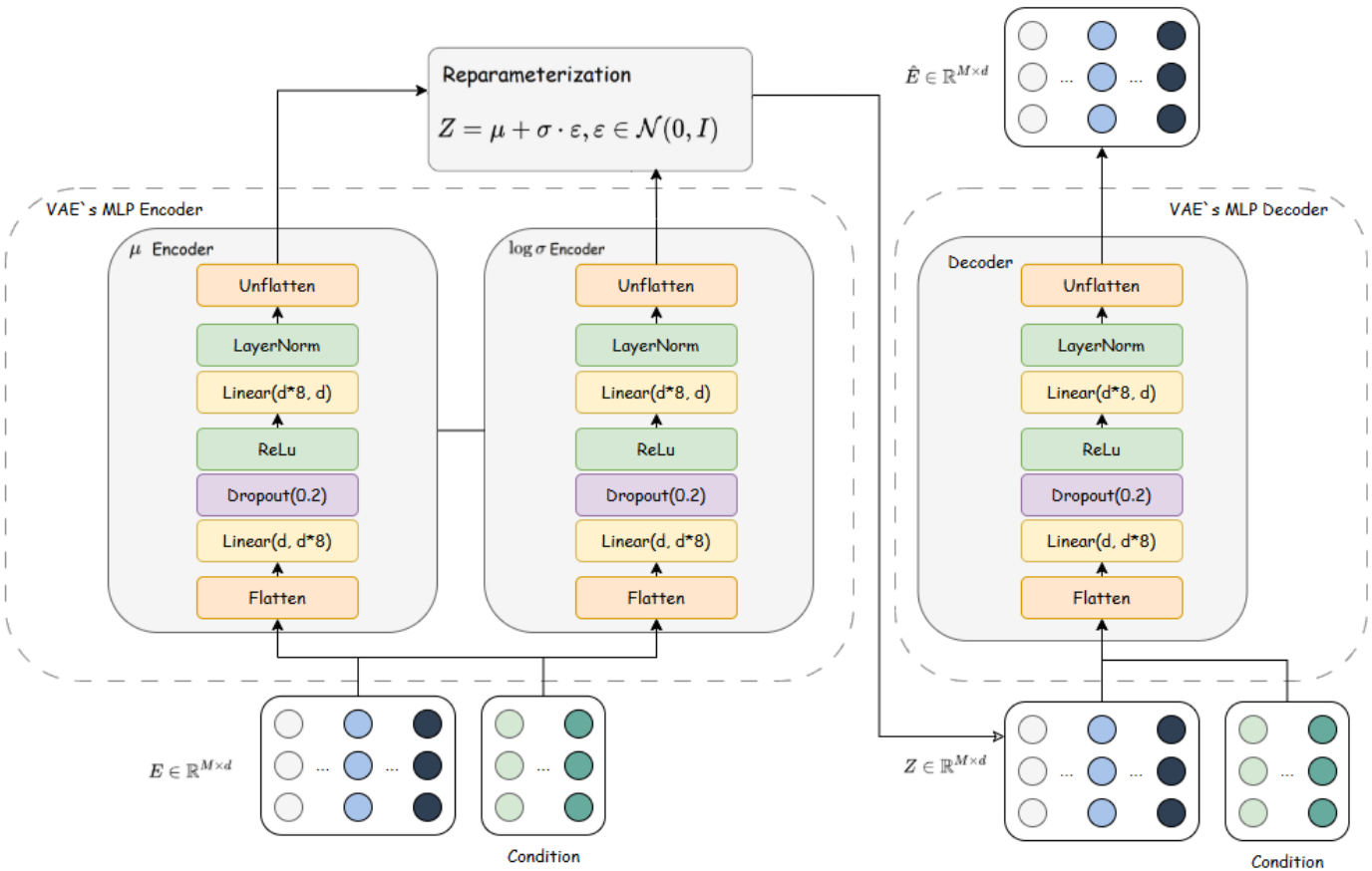
TabSyn MLP-VAE

My hypothesis is that modeling categorical features causes the greatest difficulty and the way they are encoded significantly affects the quality of the models. Since I work with tabular data, the use of transformers in VAE is unreasonable. First of all, I rewrote the VAE architecture into fully connected neural layers.



TabSyn CVAE

The second hypothesis is that VAE introduces an additional error when modeling numerical features. To test the hypothesis, I used Conditional VAE, where all numerical features serve as conditions, and VAE is trained only to encode categorical features.

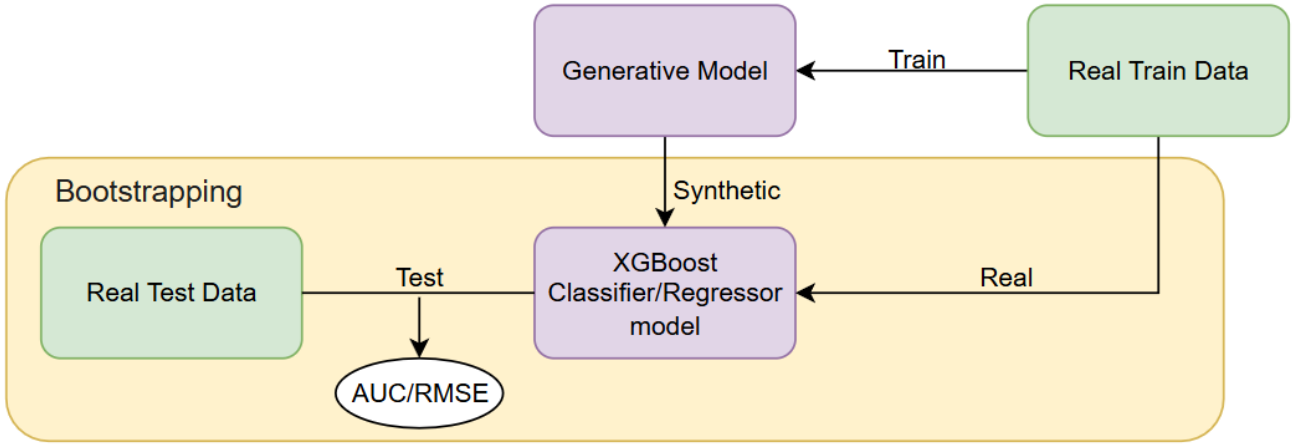


Metrics

- **Column-wise density estimation** and **Pair-wise column correlation**. These measures help to identify discrepancies between synthetic and real data at the level of basic statistical characteristics.
- **Classifier Two Sample Test** estimates how difficult it is to distinguish real data from synthetic data. This is done using the logistic regression machine learning model.
- α-**precision** and β-**recall** two sample-level metric quantifying how faithful the synthetic data is.α-precision evaluates the fidelity of synthetic data – whether each synthetic example comes from the real-data distribution, β-recall evaluates the coverage of the synthetic data.

Metrics

- **Machine Learning Efficiency** estimates the quality of a model trained on a synthetic data. The performance of synthetic data on MLE tasks is evaluated based on the divergence of test scores using the real and synthetic training data.



Datasets

I am using 5 tabular datasets from the UCI Machine Learning Repository : Adult, Default, Shoppers, Magic, Beijing, where each tabular dataset is associated with a machine learning task.

Dataset	# Rows	# Num	# Cat	# Train	# Validation	# Test	Task
Adult	48, 842	6	9	28, 943	3, 618	16, 281	C
Default	30, 000	14	11	24, 000	3, 000	3, 000	C
Shoppers	12, 330	10	8	9, 864	1, 233	1, 233	C
Magic	19, 019	10	1	15, 215	1, 902	1, 902	C
Beijing	43, 824	7	5	35, 058	4, 383	4, 383	R

Table: Statistics of datasets

Error rate (%) of **column-wise density estimation**

Method	Adult	Default	Shoppers	Magic	Beijing
SMOTE	1.60±0.23	1.48±0.15	2.68±0.19	0.91±0.05	1.85±0.21
CTGAN	16.84± 0.03	16.83±0.04	21.15±0.10	9.81±0.08	21.39±0.05
TVAE	14.22±0.08	10.17±0.05	24.51±0.06	8.25±0.06	19.16±0.06
GOGGLE[1]	16.97	17.02	22.33	1.90	16.93
GReaT[2]	12.12±0.04	19.94±0.06	14.51±0.12	16.16±0.09	8.25±0.12
STaSy	11.29±0.06	5.77±0.06	9.37±0.09	6.29±0.13	6.71±0.03
CoDi	21.38±0.06	15.77± 0.07	31.84±0.05	11.56±0.26	16.94±0.02
TabDDPM[3]	1.75±0.03	1.57± 0.08	2.72±0.13	1.01±0.09	1.30±0.03
TabSyn T-VAE	0.73±0.38	1.14±0.35	1.90±0.24	0.83±0.36	1.07±0.05
TabSyn_MLP	1.36±0.99	1.76±1.11	3.34±2.05	1.3±0.44	3.95±8.9
TabSyn_CVAE	1.78±3.8	1.96 ±1.45	2.45±1.78	1.54 ±0.45	1.33±1.64

Error rate (%) of **pair-wise column** correlation score

Method	Adult	Default	Shoppers	Magic	Beijing
SMOTE	3.28±0.29	8.41±0.38	3.56±0.22	3.16±0.41	2.39±0.35
CTGAN	20.23±1.20	26.95±0.93	13.08±0.16	7.00±0.19	22.95±0.08
TVAE	14.15±0.88	19.50±0.95	18.67±0.38	5.82±0.49	18.01±0.08
GOGGLE	45.29	21.94	23.90	9.47	45.94
GReaT	17.59±0.22	70.02±0.12	45.16±0.18	10.23±0.40	59.60±0.55
STaSy	14.51±0.25	5.96±0.26	8.49±0.15	6.61±0.53	8.00±0.10
CoDi	22.49±0.08	68.41±0.05	17.78±0.11	6.53±0.25	7.07±0.15
TabDDPM	3.01±0.25	4.89±0.10	6.61±0.16	1.70±0.22	2.71±0.09
TabSyn T-VAE	1.64±1.05	4.70±0.71	2.39±0.21	0.64±0.77	2.24±0.28
TabSyn_MLP	1.60±1.04	2.14±1.97	2.85±1.90	1.18±1.05	8.66 ±12.6
TabSyn_CVAE	5.63±6.7	11.92±19.4	3.15±2.68	2.61±3.89	4.99 ±5.43

Comparison of α-Precision scores.

Methods	**Adult**	**Default**	**Shoppers**	**Magic**	**Beijing**
CTGAN	77.74±0.15	62.08±0.08	76.97±0.39	86.90±0.22	96.27±0.14
TVAE	98.17±0.17	85.57±0.34	58.19±0.26	86.19±0.48	97.20±0.10
GOGGLE	50.68	68.89	86.95	90.88	88.81
GReaT	55.79±0.03	85.90±0.17	78.88±0.13	85.46±0.54	98.32±0.22
STaSy	82.87±0.26	90.48±0.11	89.65±0.25	86.56±0.19	89.16±0.12
CoDi	77.58±0.45	82.38±0.15	94.95±0.35	85.01±0.36	98.13±0.38
TabDDPM	96.36±0.20	97.59±0.36	88.55±0.68	98.59±0.17	97.93±0.30
TabSyn T-VAE	99.17±0.10	99.24±0.27	98.78±0.22	99.35±0.27	97.86±0.10
TabSyn_MLP	98.02±0.20	98.78±0.26	99.02±0.36	98.71±0.13	99.31±0.19
TabSyn_CVAE	98.89±0.07	94.63±0.6	99.34±0.15	91.86±0.64	98.62±0.60

Comparison of β-Recall scores.

Methods	**Adult**	**Default**	**Shoppers**	**Magic**	**Beijing**
CTGAN	30.80±0.20	18.22±0.17	31.80±0.350	11.75±0.20	34.80±0.10
TVAE	38.87±0.31	23.13±0.11	19.78±0.10	32.44±0.35	28.45±0.08
GOGGLE	8.80	14.38	9.79	9.88	19.87
GReaT	49.12±0.18	42.04±0.19	44.90±0.17	34.91±0.28	43.34±0.31
STaSy	29.21±0.34	39.31±0.39	37.24±0.45	53.97±0.57	54.79±0.18
CoDi	9.20±0.15	19.94±0.22	20.82±0.23	50.56±0.31	52.19±0.12
TabDDPM	47.05±0.25	47.83±0.35	47.79±0.25	48.46±0.42	56.92±0.13
TabSyn T-VAE	48.45±0.22	46.19±0.35	48.89±0.28	48.37±0.23	55.87±0.19
TabSyn_MLP	46.10±0.06	48.56±0.37	49.47±0.16	52.52±0.28	55.64±0.32
TabSyn_CVAE	30.97±0.84	32.78±0.44	41.99±0.33	60.12±0.15	48.75±0.64

AUC and RMSE scores of **Machine Learning Efficiency**

Methods	**Adult** AUC ↑	**Default** AUC ↑	**Shoppers** AUC ↑	**Magic** AUC ↑	**Beijing** RMSE ↓
Real	.927±.000	.770±.005	.926±.001	.946±.001	.423±.003
SMOTE	.899±.007	.741±.009	.911±.012	.934±.008	.593±.011
CTGAN	.886±.002	.696±.005	.875±.009	.855±.006	.902±.019
TVAE	.878±.004	.724±.005	.871±.006	.887±.003	.770±.011
GOGGLE	.778±.012	.584±.005	.658±.052	.654±.024	1.09±.025
GReaT	.913±.003	.755±.006	.902±.005	.888±.008	.653±.013
STaSy	.906±.001	.752±.006	.914±.005	.934±.003	.656±.014
CoDi	.871±.006	.525±.006	.865±.006	.932±.003	.818±.021
TabDDPM^2	.907±.001	.758±.004	.918±.005	.935±.003	.592±.011
TabSyn T-VAE	.909±.002	.759±.004	.917±.005	.935±.002	.582±.008
TabSyn_MLP	.908±.009	.770±.002	.916 ±.003	.939 ±.009	.583±.006
TabSyn_CVAE	.872±.001	.719±.008	.824±.003	.579±.005	.796±.002

Detection score (C2ST) using logistic regression classifier

Method	Adult	Default	Shoppers	Magic	Beijing
SMOTE	0.9710	0.9274	0.9086	0.9961	0.9888
CTGAN	0.5949	0.4875	0.7488	0.6728	0.7531
TVAE	0.6315	0.6547	0.2962	0.7706	0.8659
GOGGLE	0.1114	0.5163	0.1418	0.9526	0.4779
GReaT	0.5376	0.4710	0.4285	0.4326	0.6893
STaSy	0.4054	0.6814	0.5482	0.6939	0.7922
CoDi	0.2077	0.4595	0.2784	0.7206	0.7177
TabDDPM	0.9755	0.9712	0.8349	0.9998	0.9513
TabSyn T-VAE	0.9957	0.9674	0.9806	0.9978	0.9696
TabSyn_MLP	0.9577	0.9645	0.9073	0.9649	0.9582
TabSyn_CVAE	0.8337	0.9414	0.8698	0.9640	0.9687

Visualization of the pair-wise column correlation

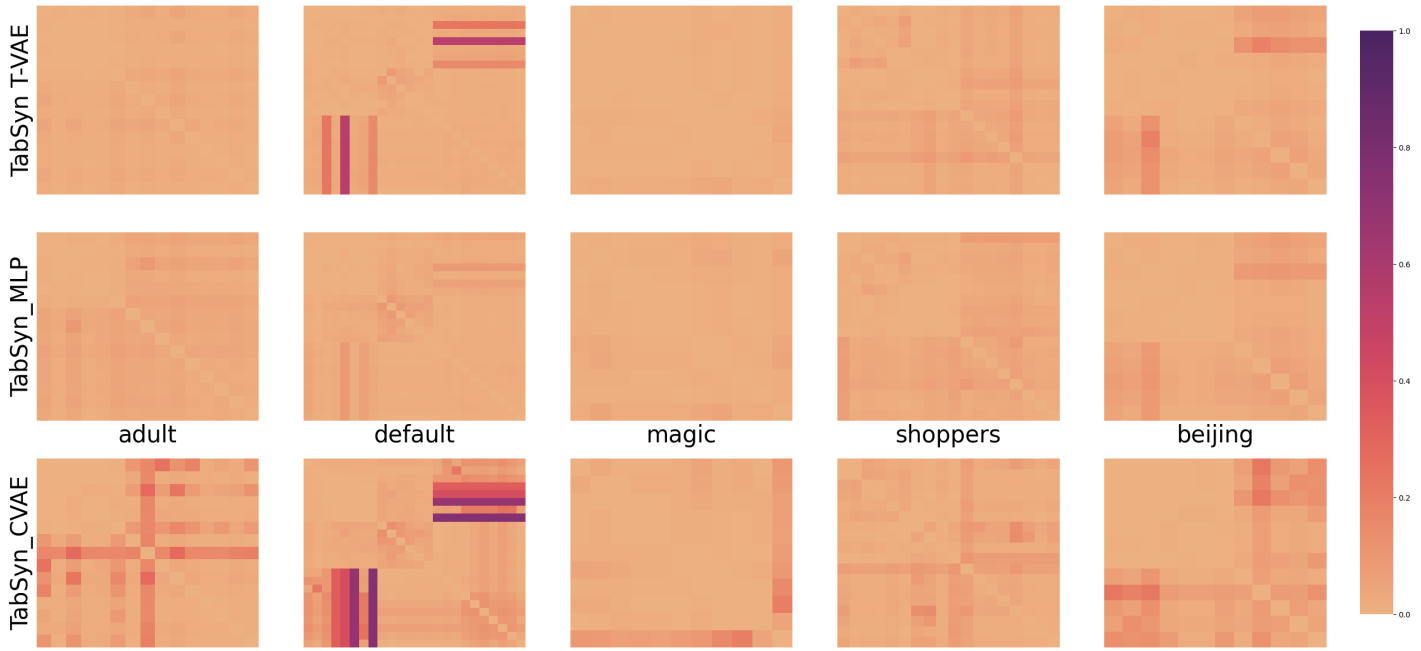


Figure: Heatmaps of the pair-wise column correlation of synthetic data v.s. the real data. The value represents the absolute divergence between the real and estimated correlations (the lighter, the better).

Visualization of the pair-wise column correlation(other research)

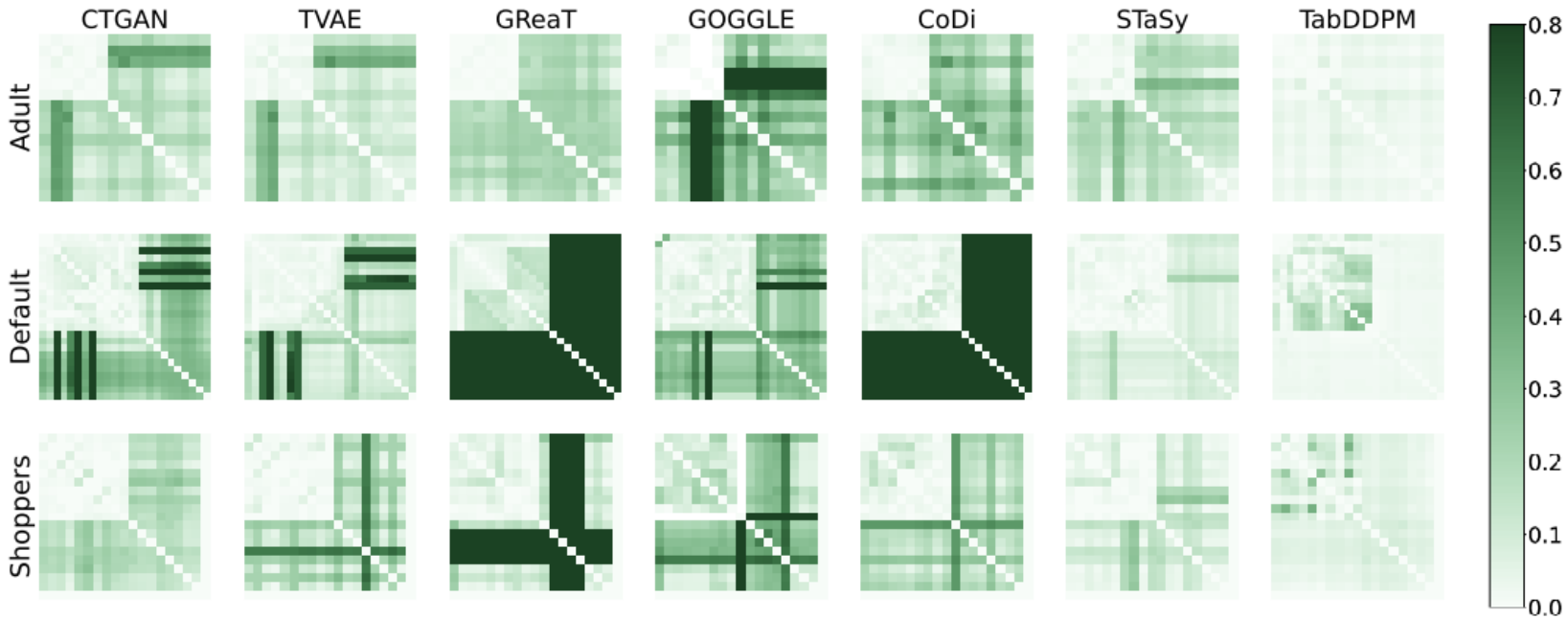
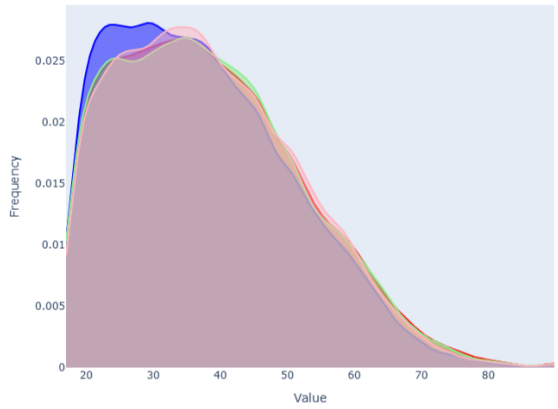
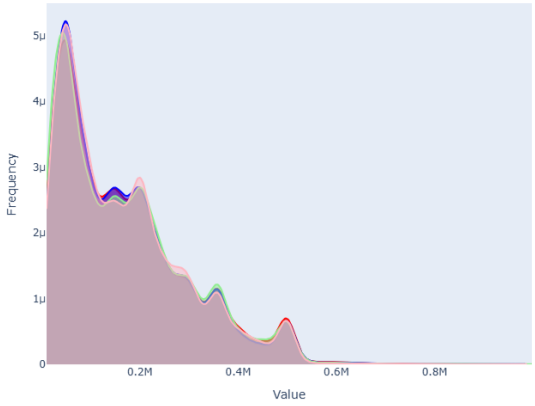
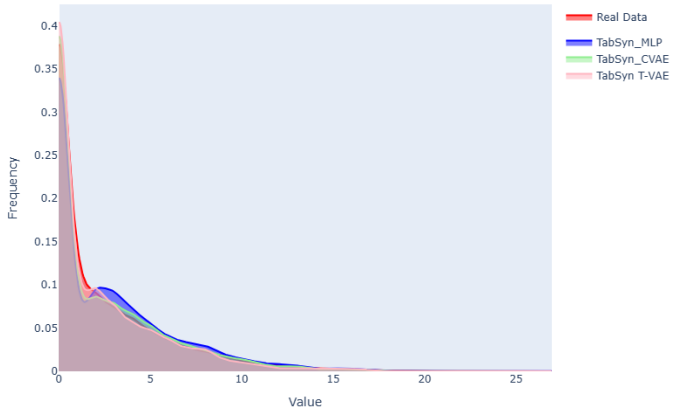


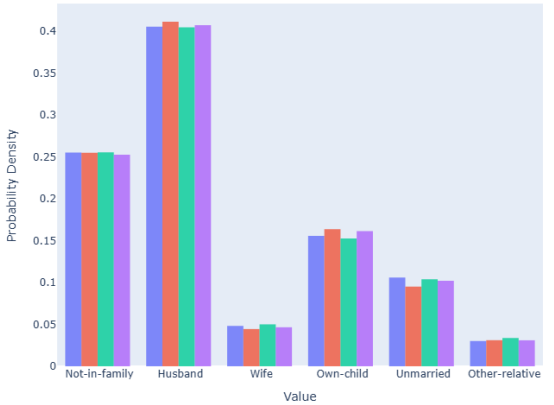
Figure: Heatmaps of the pair-wise column correlation of synthetic data v.s. the real data. The value represents the absolute divergence between the real and estimated correlations (the lighter, the better).

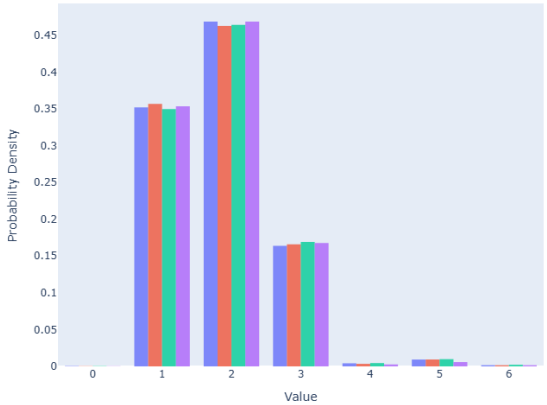
Visualization of the distribution density

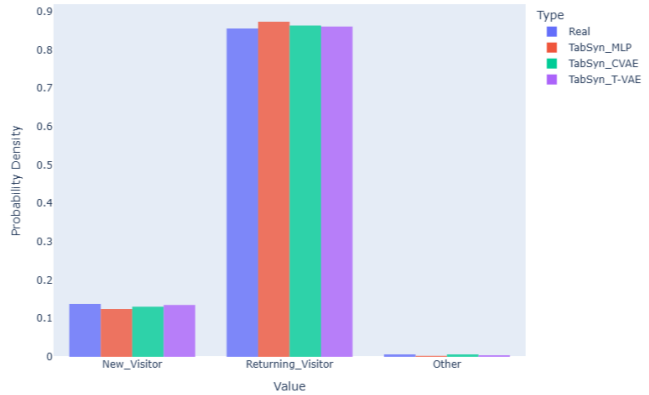












Adult

Default

Shoppers

Conclusion

- Successfully implemented and tested three different architectures
- The models were trained and tested on five popular datasets: Adult, Default, Shoppers, Magic, and Beijing, which allowed for a wide range of checks and increased confidence in the applicability of the approaches.
- The following quality indicators were used for a complex estimate of the quality of the models: MLE, C2ST, pair-wise column correlation and column-wise density estimation, α-Precision and β-Recall

All the proposed models and approaches have proven their effectiveness and practical value, which makes the work important and timely for the tasks of data analysis, protecting personal information and expanding training samples.

Q&A Session

Thank you for your attention!
Questions and discussions are welcome.

Reference

[1] Fonseca, João and Bação, Fernando. Tabular and latent space synthetic data generation: A literature review. *Journal of Big Data*, 10(1):115, 2023.
[2] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: Opportunities, challenges and pitfalls, 2021.
[3] Mikel Hernández, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022.
[4] Shuhan Zheng and Nontawat Charoenphakdee. Diffusion models for missing value imputation in tabular data. *arXiv preprint arXiv:2210.17128*, 2022.
[5] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *In Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 7335–7345, 2019.

Datasets description

- **Adult**: The "**Adult** Census Income" dataset contains the demographic and employment-related features people. The task is to predict whether an individual's income exceeds $50,000$.
- **Default**: The "**Default** of Credit Card Clients Dataset" dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The task is to predict whether the client will default payment next month.
- **Shoppers**: The "Online **Shoppers** Purchasing Intention Dataset" contains information of user's webpage visiting sessions. The task is to predict if the user's session ends with the shopping behavior.
- **Magic**: The "**Magic** Gamma Telescope" dataset is to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. The task is to classify high-energy Gamma particles in the atmosphere.
- **Beijing**: The "**Beijing** PM2.5 Data" dataset contains the hourly PM2.5 data of US Embassy in Beijing and the meteorological data from Beijing Capital International Airport. The task is to predict the PM2.5 value.