Generative Models for Incomplete Data Reconstruction

Elina Telesheva

Advisor: Mikhail Hushchyn

Modern Computer Science Master's Programme

April 2025

Motivation

- Synthesizing high-quality tabular data is important for many data science tasks:
  - privacy protection
  - dataset augmentation
  - dataset enhancement
- The generation of tabular data is challenging due to varied distributions and a mixture of data types

Problem statement

- The goal is to generate synthetic data based on real data while preserving feature statistics and making it impossible to restore real data.

Real

Name	Age
James Smith	27
Michael Johnson	15
Mary Brown	72

Synthetic

Name	Age
David Miller	19
Linda Moor	52
Richard Davis	47

Figure 1: Example of real and generated data.

Previous approaches

- **CTGAN** and **TVAE** (Xu et al. 2019) are based on GAN[a] and VAE[b] (Goodfellow et al. 2014; Kingma and Welling 2022)
- **GOOGLE** (Liu et al. 2023) is based on VAE and GNN[c]
- **GReaT** is based on auto-regressive GPT2 (Borisov et al. 2023)
- **TabDDPM**, **TabSyn**, **TabDiff**, **STaSy**, **CoDi** are diffusion models (Kotelnikov et al. 2022; Zhang et al. 2024; Shi et al. 2025; Kim et al. 2023; Lee et al. 2023)

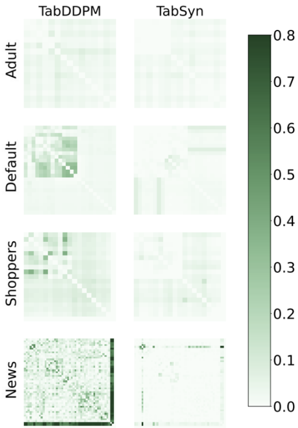

Figure 2: The value represents the absolute difference between the correlations of real and synthetic data (the lighter, the better).

[a] Generative Adversarial Network

[b] Variational autoencoder

[c] Graph neural network

TabDDPM

The focus of the research is on the TabDDPM model. It uses Gaussian diffusion to model numerical features and multinomial diffusion to model categorical features.

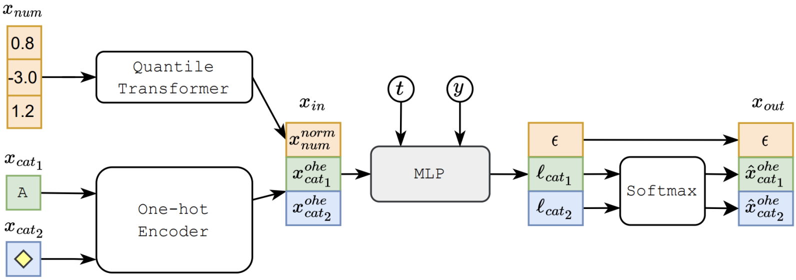

Figure 3: TabDDPM model scheme

Research question

How will the quality of the TabDDPM model change if **multinomial diffusion is excluded** from it? Let's compare the quality of the new model with the official version of TabDDPM.

Model based on TabDDPM

The multinomial diffusion part for categorical features is removed from TabDDPM model. Categorical features are processed using One-Hot Encoding and then passed to the model as numerical features.

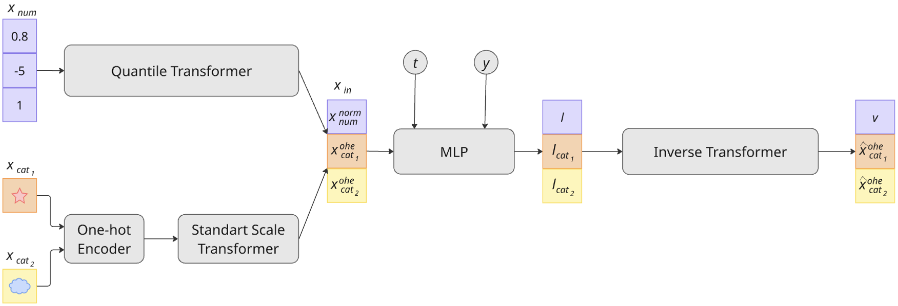

Figure 4: Considered model scheme

Metrics description

Metrics used for quality estimation of a generated dataset (Zhang et al. 2024):

- **Column-wise density estimation** and **pair-wise column correlation** estimate the density of every column and the correlation of column pairs.
- **C2ST**[1] evaluates if the synthetic data could be distinguished from the real data.
- α**-precision** and β**-recall** estimate accuracy and diversity of data.
- **DCR**[2] determines the probability that the synthetic dataset is closer to the train dataset than to the test one.
- **MLE**[3] estimates the quality of a model trained on a synthetic data.

---

[1] Classifier Two Sample Test

[2] Distance to Closest Records

[3] Machine Learning Efficiency

Metrics

Methods	Adult AUC ↑	Default AUC ↑	Shoppers AUC ↑	Magic AUC ↑	Beijing RMSE ↓	News RMSE ↓
Real	.927±.000	.770±.005	.926±.001	.946±.001	.423±.003	.842±.002
SMOTE	.899±.007	.741±.009	.911±.012	.934±.008	.593±.011	.897±.036
CTGAN	.886±.002	.696±.005	.875±.009	.855±.006	.902±.019	.880±.016
TVAE	.878±.004	.724±.005	.871±.006	.887±.003	.770±.011	1.01±.016
GOGGLE	.778±.012	.584±.005	.658±.052	.654±.024	1.09±.025	.877±.002
GReaT	.913±.003	.755±.006	.902±.005	.888±.008	.653±.013	OOM
STaSy	.906±.001	.752±.006	.914±.005	.934±.003	.656±.014	.871±.002
CoDi	.871±.006	.525±.006	.865±.006	.932±.003	.818±.021	1.21±.005
TabSyn	.915±.002	.764±.004	.920±.005	.938±.002	.582±.008	.861±.027
TabDDPM	.904±.006	.758±.013	.916±.003	.927±.004	.592±.011	4.86±3.04
RESEARCH	.873±0.006	.715±0.014	.920±0.005	.925±0.005	*XXX*	*XXX*

Table 1: AUC (classification task) and RMSE (regression task) scores of Machine Learning Efficiency

Pair-wise column correlation for ‘Adult’

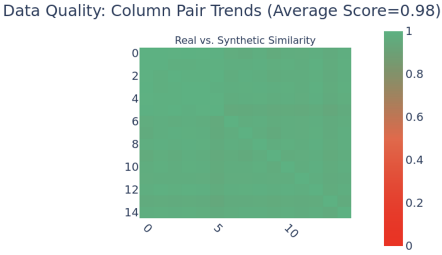

Figure 5: TabDDPM

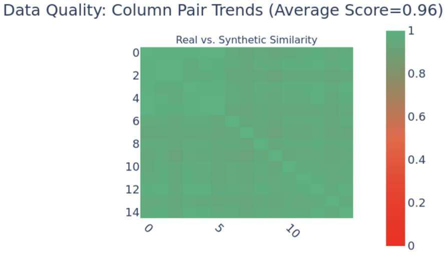

Figure 6: Our model

Conclusion

- Developed a **new method** to generate synthetic data.
- Removing the multinomial diffusion model from TabDDPM allows us to **simplify the model while maintaining similar generation quality**.
- Tested **5 different versions** of the model, and went through more than **15 different combinations** of hyperparameter settings for each version.
  - There is an **impact of noise** level on the quality of generation
- Studied and **calculated the quality metrics**.
- Figured out how the TabDDPM model works and **tested it on several datasets**.
- See more results in Appendix and GitHub.

References

Borisov, V., K. Seßler, T. Leemann, M. Pawelczyk, and G. Kasneci (2023). Language models are realistic tabular data generators.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative adversarial networks.

Kim, J., C. Lee, and N. Park (2023). Stasy: Score-based tabular data synthesis.

Kingma, D. P. and M. Welling (2022). Auto-encoding variational bayes.

Kotelnikov, A., D. Baranchuk, I. Rubachev, and A. Babenko (2022). Tabddpm: Modelling tabular data with diffusion models.

Lee, C., J. Kim, and N. Park (2023). Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis.

Liu, T., Z. Qian, J. Berrevoets, and M. van der Schaar (2023). Goggle: Generative modelling for tabular data by learning relational structure.

Shi, J., M. Xu, H. Hua, H. Zhang, S. Ermon, and J. Leskovec (2025). Tabdiff: a mixed-type diffusion model for tabular data generation.

Xu, L., M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni (2019). Modeling tabular data using conditional gan.

Zhang, H., J. Zhang, B. Srinivasan, Z. Shen, X. Qin, C. Faloutsos, H. Rangwala, and G. Karypis (2024). Mixed-type tabular data synthesis with score-based diffusion in latent space.

Appendix

Datasets description (1/2)

- **Adult**: The 'Adult Census Income' dataset contains information about demographic, educational and employment characteristics of individuals. The task is to predict whether the annual income of a person exceeds $50,000.
- **Default**: The 'Default of Credit Card Clients' dataset contains data of customers' default payments in Taiwan. The task is to predict whether a client will default payment next month.
- **Shoppers**: The 'Online Shoppers Purchasing Intention Dataset' contains information about a person's web browsing activity. The task is to predict whether a session will result in shopping behavior.

Datasets description (2/2)

- **Magic**: The 'MAGIC Gamma Telescope' dataset contains data of the simulation of high-energy gamma-ray particles using the Cherenkov ground-based telescope. The goal is to classify these high-energy particles in the atmosphere.
- **Beijing**: The 'Beijing PM2.5' dataset provides hourly PM2.5 data of US Embassy in Beijing, as well as meteorological data collected at Beijing Capital International Airport. The task is to use this data to predict PM2.5 levels.
- **News**: The 'Online News Popularity' dataset contains a variety of features about articles published by Mashable over a two-year period. The task is to predict the number of shares in social networks.

Metrics

Method	Adult	Default	Shoppers	Magic	Beijing	News
SMOTE	1.60±0.23	1.48±0.15	2.68±0.19	0.91±0.05	1.85±0.21	5.31±0.46
CTGAN	16.84±0.03	16.83±0.04	21.15±0.10	9.81±0.08	21.39±0.05	16.09±0.02
TVAE	14.22±0.08	10.17±0.05	24.51±0.06	8.25±0.06	19.16±0.06	16.62±0.03
GOGGLE	16.97	17.02	22.33	1.90	16.93	25.32
GReaT	12.12±0.04	19.94±0.06	14.51±0.12	16.16±0.09	8.25±0.12	OOM
STaSy	11.29±0.06	5.77±0.06	9.37±0.09	6.29±0.03	6.71±0.03	6.89±0.03
CoDi	21.38±0.06	15.77±0.06	31.84±0.05	11.56±0.26	16.94±0.02	32.27±0.04
TabSyn	0.58±0.06	0.85±0.04	1.43±0.24	0.88±0.09	1.12±0.05	1.64±0.04
TabDDPM	1.18±0.66	1.45±0.58	2.92±1.10	1.040±0.40	1.30±0.03	78.75±0.01
RESEARCH	1.57±0.98	1.42±0.57	1.59±0.76	0.905±0.29	*XXX*	*XXX*

Table 2: Error rate (%) of column-wise density estimation

Metrics

Method	Adult	Default	Shoppers	Magic	Beijing	News
SMOTE	3.28±0.29	8.41±0.38	3.56±0.22	3.16±0.41	2.39±0.35	5.38±0.76
CTGAN	20.23±1.20	26.95±0.93	13.08±0.16	7.00±0.19	22.95±0.08	5.37±0.05
TVAE	14.15±0.88	19.50±0.95	18.67±0.38	5.82±0.49	18.01±0.08	6.17±0.09
GOGGLE	45.29	21.94	23.90	9.47	45.94	23.19
GReaT	17.59±0.22	70.02±0.12	45.16±0.18	10.23±0.40	59.60±0.55	OOM
STaSy	14.51±0.25	5.96±0.26	8.49±0.15	6.61±0.53	8.00±0.10	3.07±0.04
CoDi	22.49±0.08	68.41±0.05	17.78±0.11	6.53±0.25	7.07±0.15	11.10±0.01
TabSyn	1.54±0.27	2.05±0.12	2.07±0.21	1.06±0.31	2.24±0.28	1.44±0.03
TabDDPM	2.16±1.11	2.91±2.94	9.06±8.33	0.746±0.76	2.71±0.09	13.16±0.11
RESEARCH	3.00±1.59	2.21±1.38	2.39±1.39	1.15±2.13	*XXX*	*XXX*

Table 3: Error rate (%) of pair-wise column correlation score

Metrics

Methods	Adult AUC ↑	Default AUC ↑	Shoppers AUC ↑	Magic AUC ↑	Beijing RMSE ↓	News RMSE ↓
Real	.927±.000	.770±.005	.926±.001	.946±.001	.423±.003	.842±.002
SMOTE	.899±.007	.741±.009	.911±.012	.934±.008	.593±.011	.897±.036
CTGAN	.886±.002	.696±.005	.875±.009	.855±.006	.902±.019	.880±.016
TVAE	.878±.004	.724±.005	.871±.006	.887±.003	.770±.011	1.01±.016
GOGGLE	.778±.012	.584±.005	.658±.052	.654±.024	1.09±.025	.877±.002
GReaT	.913±.003	.755±.006	.902±.005	.888±.008	.653±.013	OOM
STaSy	.906±.001	.752±.006	.914±.005	.934±.003	.656±.014	.871±.002
CoDi	.871±.006	.525±.006	.865±.006	.932±.003	.818±.021	1.21±.005
TabSyn	.915±.002	.764±.004	.920±.005	.938±.002	.582±.008	.861±.027
TabDDPM	.904±.006	.767±.008	.916±.003	.927±.004	.592±.011	4.86±3.04
Research	.873±0.006	.765±0.008	.920±0.005	.925±0.005	*XXX*	*XXX*

Table 4: AUC (classification task) and RMSE (regression task) scores of Machine Learning Efficiency

Metrics

Methods	Adult	Default	Shoppers	Magic	Beijing	News
CTGAN	77.74±0.15	62.08±0.08	76.97±0.39	86.90±0.22	96.27±0.14	96.96±0.17
TVAE	98.17±0.17	85.57±0.34	58.19±0.26	86.19±0.48	97.20±0.10	86.41±0.17
GOGGLE	50.68	68.89	86.95	90.88	88.81	86.41
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	94.76±0.33
CoDi	77.58±0.45	82.38±0.15	94.95±0.35	85.01±0.36	98.13±0.38	87.15±0.12
TabSyn	99.52±0.10	99.26±0.27	99.16±0.22	99.38±0.27	98.47±0.10	96.80±0.25
TabDDPM	95.15±0.20	97.76±0.36	95.14±0.68	98.11±0.17	97.93±0.30	0.00±0.00
RESEARCH	94.61±0.22	97.63±0.29	95.37±0.64	99.41±0.51	*XXX*	*XXX*

Table 5: Comparison of α-Precision scores

Metrics

Methods	Adult	Default	Shoppers	Magic	Beijing	News
CTGAN	30.80±0.20	18.22±0.17	31.80±0.35	11.75±0.20	34.80±0.10	24.97±0.29
TVAE	38.87±0.31	23.13±0.11	19.78±0.10	32.44±0.35	28.45±0.08	29.66±0.21
GOGGLE	8.80	14.38	9.79	9.88	19.87	2.03
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	39.42±0.32
CoDi	9.20±0.15	19.94±0.22	20.82±0.23	50.56±0.31	52.19±0.12	34.40±0.31
TabSyn	47.56±0.22	48.00±0.35	48.95±0.28	48.03±0.23	55.84±0.19	45.04±0.34
TabDDPM	49.76±0.25	47.25±0.35	49.76±0.25	47.96±0.42	56.92±0.13	0.00±0.00
Research	48.28±0.25	46.69±0.31	51.06±0.30	49.36±0.12	*XXX*	*XXX*

Table 6: Comparison of β-Recall scores

Metrics

Method	Adult	Default	Shoppers	Magic	Beijing	News
SMOTE	0.9710	0.9274	0.9086	0.9961	0.9888	0.9344
CTGAN	0.5949	0.4875	0.7488	0.6728	0.7531	0.6947
TVAE	0.6315	0.6547	0.2962	0.7706	0.8659	0.4076
GOGGLE	0.1114	0.5163	0.1418	0.9526	0.4779	0.0745
GReaT	0.5376	0.4710	0.4285	0.4326	0.6893	–
STaSy	0.4054	0.6814	0.5482	0.6939	0.7922	0.5287
CoDi	0.2077	0.4595	0.2784	0.7206	0.7177	0.0201
TabSyn	0.9986	0.9870	0.9740	0.9732	0.9603	0.9749
TabDDPM	0.9530	0.9834	0.8666	0.9998	0.9513	0.0002
RESEARCH	0.9350	0.9653	0.9300	0.9998	*XXX*	*XXX*

Table 7: Detection score (C2ST) using logistic regression classifier. Higher scores indicate better performance.

Appendix

Method	Default	Shoppers
SMOTE	91.41%±3.42	96.40%±4.70
STaSy	50.23%±0.09	51.53%±0.16
CoDi	51.82%±0.26	51.06%±0.18
TabSyn	51.20%±0.18	52.90%±0.22
TabDDPM	50.17%±1.20	52.42%±0.25
RESEARCH	51.59%±1.26	51.07%±1.01

Table 8: DCR score