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 (Kotelnikov et al. 2022; Zhang et al. 2024; Shi et al. 2025)

TabDDPM

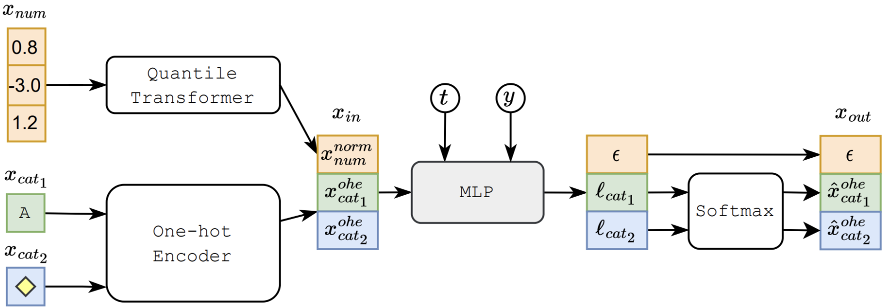

Figure 1: TabDDPM model scheme

Difficult to generate categorical features

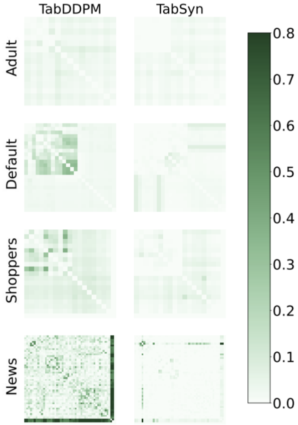

Figure 2: The value represents the absolute difference between the correlations of real and synthetic data (the lighter, the better).

Research question

Is it possible to maintain the quality of data generation without using multinomial diffusion model based on TabDDPM?

Model based on TabDDPM

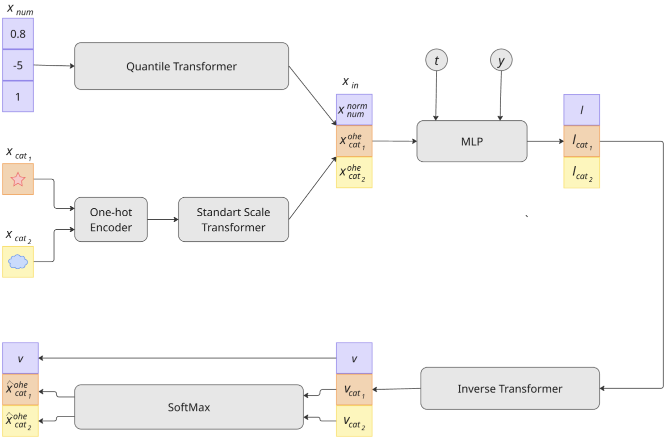

Figure 3: Considered model

Data Similarity

Methods	Adult	Default	Shoppers	Magic	Beijing	News
TabDDPM	$1.75_{\pm 0.03}$	$1.57_{\pm 0.08}$	$2.72_{\pm 0.13}$	$1.01_{\pm 0.09}$	$1.30_{\pm 0.03}$	$78.75_{\pm 0.01}$
TabSyn	$0.58_{\pm 0.06}$	$0.85_{\pm 0.04}$	$1.43_{\pm 0.24}$	$0.88_{\pm 0.09}$	$1.12_{\pm 0.05}$	$1.64_{\pm 0.04}$
Research	$1.866_{\pm 1.056}$	*XXX*	*XXX*	*XXX*	*XXX*	*XXX*

Table 1: Error rate (%) of column-wise density estimation

Methods	Adult	Default	Shoppers	Magic	Beijing	News
TabDDPM	$3.01_{\pm 0.25}$	$4.89_{\pm 0.1}$	$6.61_{\pm 0.16}$	$1.70_{\pm 0.22}$	$2.71_{\pm 0.09}$	$13.16_{\pm 0.11}$
TabSyn	$1.54_{\pm 0.27}$	$2.05_{\pm 0.12}$	$2.07_{\pm 0.21}$	$1.06_{\pm 0.31}$	$2.24_{\pm 0.28}$	$1.44_{\pm 0.03}$
Research	$3.418_{\pm 1.810}$	*XXX*	*XXX*	*XXX*	*XXX*	*XXX*

Table 2: Error rate (%) of pair-wise column correlation score

Machine Learning Efficiency

Methods	AUC ↑				RMSE ↓	
	Adult	Default	Shoppers	Magic	Beijing	News
Real	$.927_{\pm .000}$	$.770_{\pm .005}$	$.926_{\pm .001}$	$.946_{\pm .001}$	$.423_{\pm .003}$	$.842_{\pm .002}$
TabDDPM	$.907_{\pm .001}$	$.758_{\pm .004}$	$.918_{\pm .005}$	$.935_{\pm .003}$	$.592_{\pm .011}$	$4.86_{\pm 3.04}$
TabSyn	$.915_{\pm .002}$	$.764_{\pm .004}$	$.920_{\pm .005}$	$.938_{\pm .002}$	$.582_{\pm .008}$	$.861_{\pm .027}$
Research	$.875_{\pm 0.008}$	*XXX*	*XXX*	*XXX*	*XXX*	*XXX*

Table 3: Results of AUC (classification task) and RMSE (regression task) scores.

Conclusion

- Removing the multinomial diffusion model from TabDDPM allows us to simplify the model without a significant quality loss.

References

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

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

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.