TabDDPM: Modelling Tabular Data with Diffusion Models

Akim Kotelnikov
HSE, Yandex

Dmitry Baranchuk
Yandex

Ivan Rubachev
HSE, Yandex

Artem Babenko
Yandex

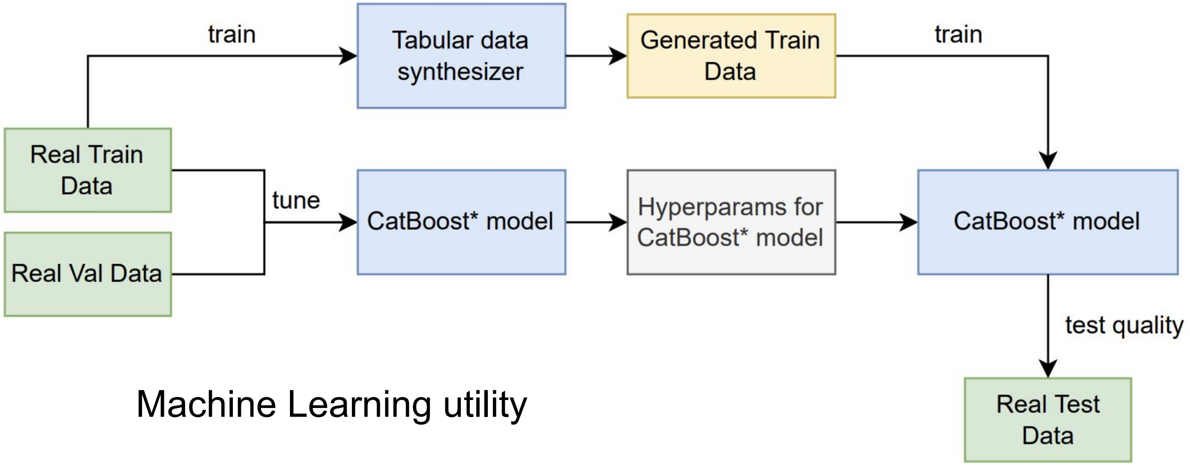
Problem statement. Evaluation

Why?

- In addition to real data (like augmentations)
- Privacy-oriented tasks

Evaluation:

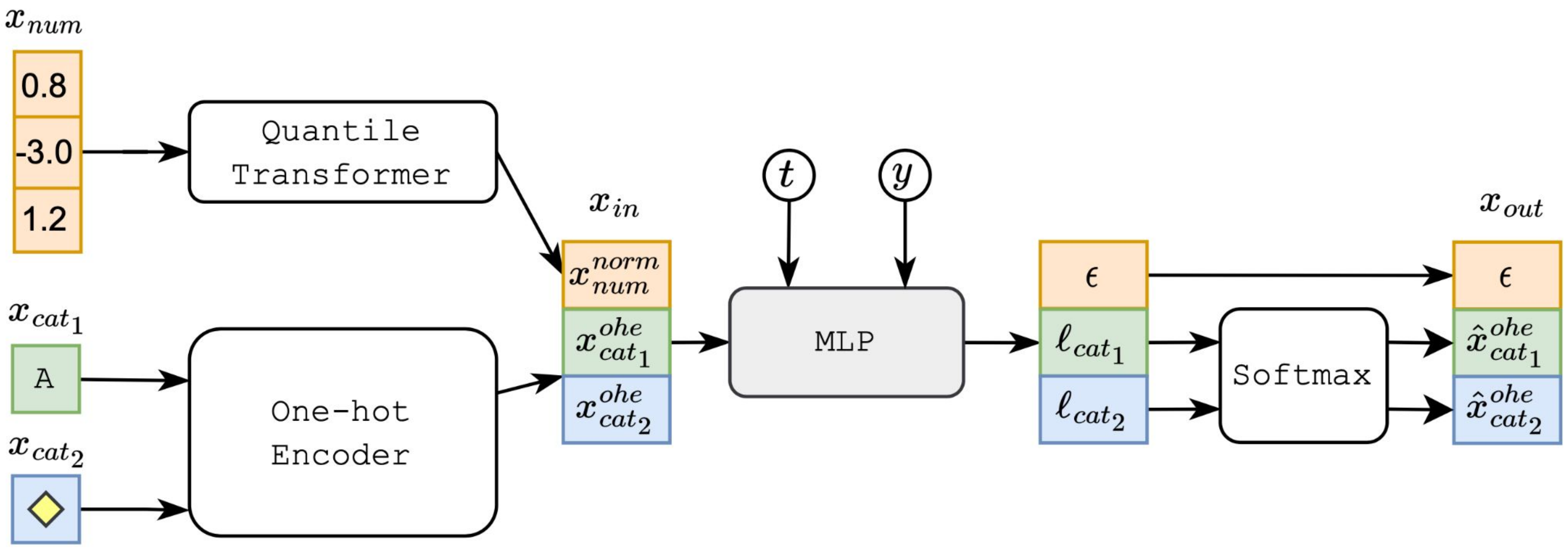
- Computer Vision: IS, FID, human eval, diversity metrics
- Tabular Data: no benchmarks, many different datasets, almost no human eval



Machine Learning utility

TabDDPM scheme

- Gaussian diffusion for numerical features (simple MSE loss)
- Multinomial diffusion for categorical features (VLB)
- Conditional model for classification problems, joint for regression
- A simple MLP model approximates the reverse process



Multinomial diffusion

Multinomial diffusion models (Hoogeboom et al., 2021) are designed to generate categorical data where $x_t \in \{0,1\}^K$ is a one-hot encoded categorical variable with K values. The multinomial forward diffusion process defines $q\left(x_t|x_{t-1}\right)$ as a categorical distribution that corrupts the data by uniform noise over K classes:

$$
\begin{aligned}
&q(x_t|x_{t-1}) := Cat\left(x_t; (1-\beta_t)\, x_{t-1} + \beta_t/K\right) \\
&q\left(x_T\right) := Cat\left(x_T; 1/K\right) \\
&q\left(x_t|x_0\right) = Cat\left(x_t; \bar{\alpha}_t x_0 + \left(1-\bar{\alpha}_t\right)/K\right)
\end{aligned}
$$

From the equations above, the posterior $q(x_{t-1}|x_t, x_0)$ can be derived:

$$
q\left(x_{t-1}|x_t, x_0\right) = Cat\left(x_{t-1}; \pi / \sum_{k=1}^{K} \pi_k\right)
$$

where $\pi = [\alpha_t x_t + (1-\alpha_t)/K] \odot [\bar{\alpha}_{t-1} x_0 + (1-\bar{\alpha}_{t-1})/K]$.

The reverse distribution $p_\theta\left(x_{t-1}|x_t\right)$ is parameterized as $q\left(x_{t-1}|x_t, \hat{x}_0(x_t, t)\right)$, where $\hat{x}_0$ is predicted by a neural network. Then, the model is trained to maximize the variational lower bound (1).

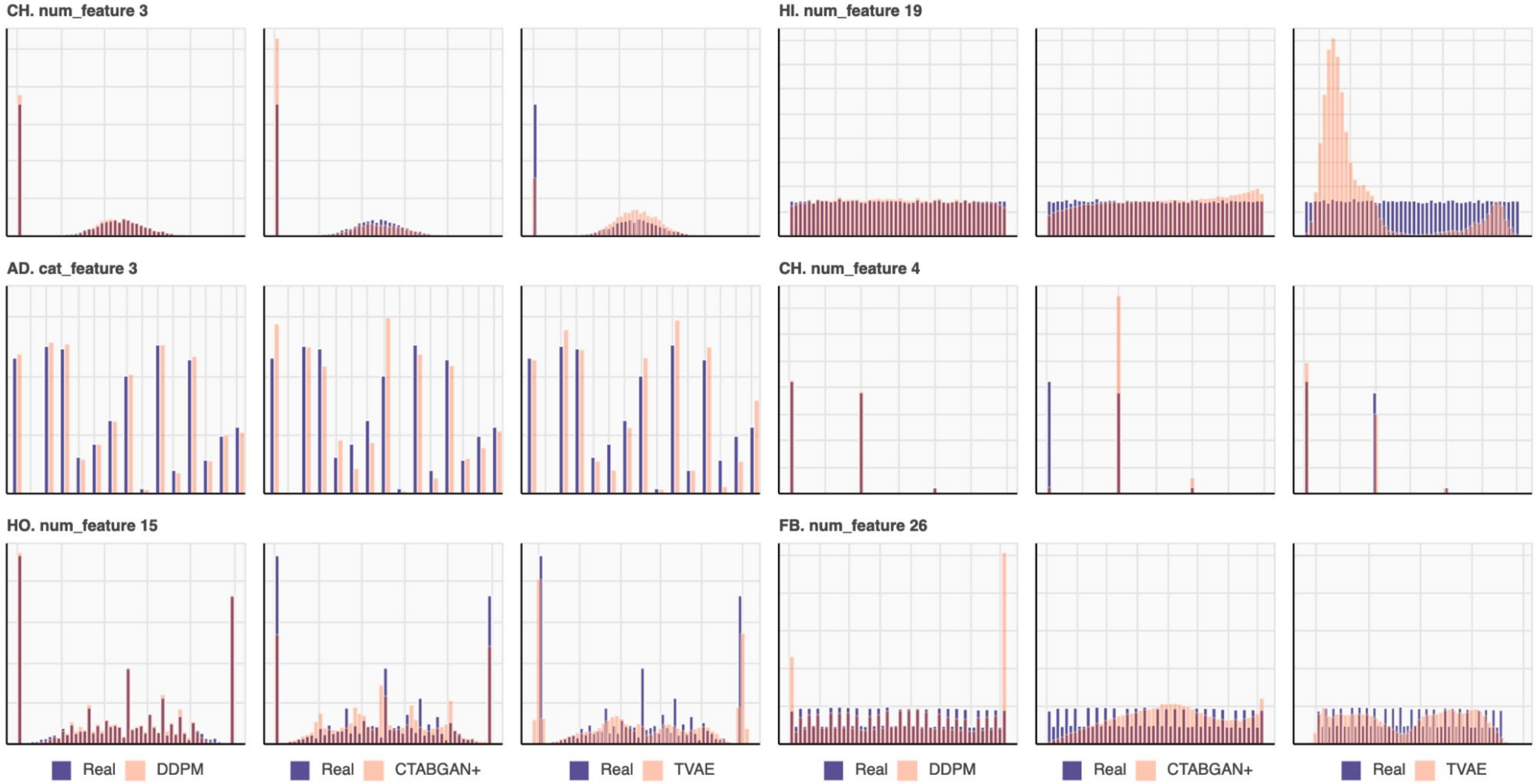
Note on tuning

- Using the Optuna library, we tune hyperparameters of TabDDPM
- We use 50 tuning iterations and validation set to calculate ML utility

Hyperparameter	Search space
Learning rate	$\mathrm{LogUniform}[0.00001, 0.003]$
Batch size	$\mathrm{Cat}\{256, 4096\}$
Diffusion timesteps	$\mathrm{Cat}\{100, 1000\}$
Training iterations	$\mathrm{Cat}\{5000, 10000, 20000\}$
# MLP layers	$\mathrm{Int}\{2, 4, 6, 8\}$
MLP width of layers	$\mathrm{Int}\{128, 256, 512, 1024\}$
Proportion of samples	$\mathrm{Float}\{0.25, 0.5, 1, 2, 4, 8\}$
Dropout	0.0
Scheduler	cosine (Nichol, 2021)
Gaussian diffusion loss	MSE
Number of tuning trials	50

Prior methods. Visualization of features

- TVAE [1], CTABGAN [2], CTABGAN+ [3]



ML utility with CatBoost models

	AB ($R2$)	AD ($F1$)	BU ($F1$)	CA ($R2$)	CAR ($F1$)	CH ($F1$)	DE ($F1$)	DI ($F1$)
TVAE	$0.433 \pm .008$	$0.781 \pm .002$	$0.864 \pm .005$	$0.752 \pm .001$	$0.717 \pm .001$	$0.732 \pm .006$	$0.656 \pm .007$	$\mathbf{0.714 \pm .039}$
CTABGAN	–	$0.783 \pm .002$	$0.855 \pm .005$	–	$0.717 \pm .001$	$0.688 \pm .006$	$0.644 \pm .011$	$\mathbf{0.731 \pm .022}$
CTABGAN+	$0.467 \pm .004$	$0.772 \pm .003$	$0.884 \pm .005$	$0.525 \pm .004$	$0.733 \pm .001$	$0.702 \pm .012$	$0.686 \pm .004$	$\mathbf{0.734 \pm .020}$
SMOTE	$\mathbf{0.549 \pm .005}$	$0.791 \pm .002$	$0.891 \pm .003$	$\mathbf{0.840 \pm .001}$	$0.732 \pm .001$	$0.743 \pm .005$	$\mathbf{0.693 \pm .003}$	$0.683 \pm .037$
TabDDPM	$\mathbf{0.550 \pm .010}$	$\mathbf{0.795 \pm .001}$	$\mathbf{0.906 \pm .003}$	$0.836 \pm .002$	$\mathbf{0.737 \pm .001}$	$\mathbf{0.755 \pm .006}$	$\mathbf{0.691 \pm .004}$	$\mathbf{0.740 \pm .020}$
Real	$0.556 \pm .004$	$0.815 \pm .002$	$0.906 \pm .002$	$0.857 \pm .001$	$0.738 \pm .001$	$0.740 \pm .009$	$0.688 \pm .003$	$0.785 \pm .013$

	FB ($R2$)	GE ($F1$)	HI ($F1$)	HO ($R2$)	IN ($R2$)	KI ($R2$)	MI ($F1$)	WI ($F1$)
TVAE	$0.685 \pm .003$	$0.434 \pm .006$	$0.638 \pm .003$	$0.493 \pm .006$	$0.784 \pm .010$	$0.824 \pm .003$	$0.912 \pm .001$	$0.501 \pm .012$
CTABGAN	–	$0.392 \pm .006$	$0.575 \pm .004$	–	–	–	$0.889 \pm .002$	$\mathbf{0.906 \pm .019}$
CTABGAN+	$0.509 \pm .011$	$0.406 \pm .009$	$0.664 \pm .002$	$0.504 \pm .005$	$0.797 \pm .005$	$0.444 \pm .014$	$0.892 \pm .002$	$0.798 \pm .021$
SMOTE	$\mathbf{0.803 \pm .002}$	$\mathbf{0.658 \pm .007}$	$\mathbf{0.722 \pm .001}$	$0.662 \pm .004$	$\mathbf{0.812 \pm .002}$	$\mathbf{0.842 \pm .004}$	$0.932 \pm .001$	$\mathbf{0.913 \pm .007}$
TabDDPM	$0.713 \pm .002$	$0.597 \pm .006$	$\mathbf{0.722 \pm .001}$	$\mathbf{0.677 \pm .010}$	$0.809 \pm .002$	$\mathbf{0.833 \pm .014}$	$\mathbf{0.936 \pm .001}$	$\mathbf{0.904 \pm .009}$
Real	$0.837 \pm .001$	$0.636 \pm .007$	$0.724 \pm .001$	$0.662 \pm .003$	$0.814 \pm .001$	$0.907 \pm .002$	$0.934 \pm .000$	$0.898 \pm .006$

ML utility with Simple models

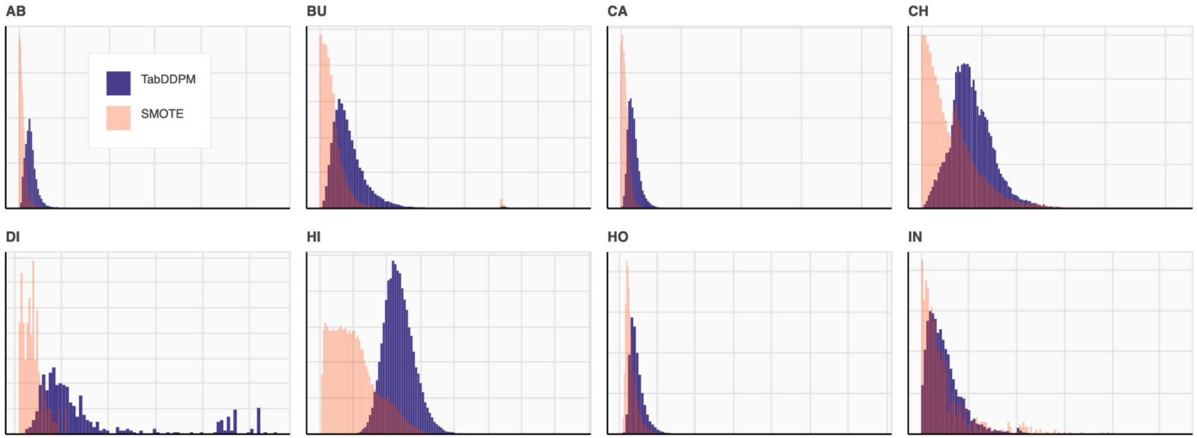
	AB ($R2$)	AD ($F1$)	BU ($F1$)	CA ($R2$)	CAR ($F1$)	CH ($F1$)	DE ($F1$)	DI ($F1$)
TVAE	$0.238\pm.012$	$0.742\pm.001$	$0.779\pm.004$	-13.0 ± 1.51	$0.693\pm.002$	$0.684\pm.003$	$0.643\pm.003$	$0.712\pm.010$
CTABGAN	–	$0.737\pm.007$	$0.786\pm.008$	–	$0.684\pm.003$	$0.636\pm.010$	$0.614\pm.007$	$0.655\pm.015$
CTABGAN+	$0.316\pm.024$	$0.730\pm.007$	$0.837\pm.006$	$-7.59\pm.645$	$\mathbf{0.708\pm.002}$	$0.650\pm.008$	$0.648\pm.008$	$\mathbf{0.727\pm.023}$
SMOTE	$\mathbf{0.400\pm.009}$	$0.750\pm.004$	$0.842\pm.003$	$0.667\pm.006$	$0.693\pm.001$	$0.690\pm.003$	$0.649\pm.003$	$0.677\pm.013$
TabDDPM	$\mathbf{0.392\pm.009}$	$\mathbf{0.758\pm.005}$	$\mathbf{0.851\pm.003}$	$\mathbf{0.695\pm.002}$	$0.696\pm.001$	$\mathbf{0.693\pm.003}$	$\mathbf{0.659\pm.003}$	$0.675\pm.011$
Real	$0.423\pm.009$	$0.750\pm.006$	$0.845\pm.004$	$0.663\pm.002$	$0.683\pm.002$	$0.679\pm.003$	$0.648\pm.003$	$0.699\pm.012$

	FB ($R2$)	GE ($F1$)	HI ($F1$)	HO ($R2$)	IN ($R2$)	KI ($R2$)	MI ($F1$)	WI ($F1$)
TVAE	$\ll 0$	$0.372\pm.006$	$0.590\pm.004$	$0.174\pm.012$	$0.470\pm.024$	$0.666\pm.006$	$\mathbf{0.880\pm.002}$	$0.497\pm.001$
CTABGAN	–	$0.339\pm.009$	$0.539\pm.006$	–	–	–	$0.856\pm.003$	$0.656\pm.011$
CTABGAN+	$\ll 0$	$0.373\pm.009$	$0.598\pm.004$	$0.222\pm.042$	$0.669\pm.018$	$0.197\pm.051$	$0.867\pm.002$	$0.653\pm.027$
SMOTE	$\mathbf{0.651\pm.002}$	$\mathbf{0.478\pm.005}$	$0.664\pm.003$	$0.394\pm.006$	$0.709\pm.008$	$\mathbf{0.751\pm.005}$	$0.860\pm.001$	$\mathbf{0.793\pm.004}$
TabDDPM	$0.527\pm.005$	$0.462\pm.005$	$\mathbf{0.670\pm.002}$	$\mathbf{0.426\pm.007}$	$\mathbf{0.734\pm.007}$	$0.611\pm.013$	$0.850\pm.004$	$\mathbf{0.792\pm.004}$
Real	$0.645\pm.005$	$0.431\pm.005$	$0.663\pm.002$	$0.415\pm.007$	$0.708\pm.007$	$0.768\pm.013$	$0.850\pm.004$	$0.684\pm.004$

Privacy. Distance to closest record

- For each synthetic sample, we find the minimum distance to real datapoints and take the median of these distances
- Low DCR values indicate that all synthetic samples are essentially copies of some real datapoints
- Larger DCR values indicate that the generative model can produce something “new” rather than just copies of real data

TabDDPM vs SMOTE



	AB		AD		BU		CA		CAR		CH		DE		DI	
	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR
SMOTE	0.549	0.014	0.791	0.024	0.891	0.054	0.840	0.014	0.732	0.007	0.743	0.077	0.693	0.027	0.683	0.068
TabDDPM	0.550	**0.050**	0.795	**0.104**	0.906	**0.143**	0.836	**0.041**	0.737	**0.012**	0.755	**0.157**	0.691	**0.112**	0.740	**0.204**

	FB		GE		HI		HO		IN		KI		MI		WI	
	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR
SMOTE	0.803	0.027	0.658	0.023	0.722	0.319	0.662	0.056	0.812	0.030	0.842	0.066	0.932	0.016	0.913	0.007
TabDDPM	0.713	**0.112**	0.597	**0.059**	0.722	**0.449**	0.677	**0.086**	0.809	**0.041**	0.833	**0.189**	0.936	**0.022**	0.904	**0.016**

References

- [1] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. 2019
- [2] Ctab-gan: Effective table data synthesizing, PMLR 2021
- [3] Ctab-gan+: Enhancing tabular data synthesis. arXiv preprint 2022