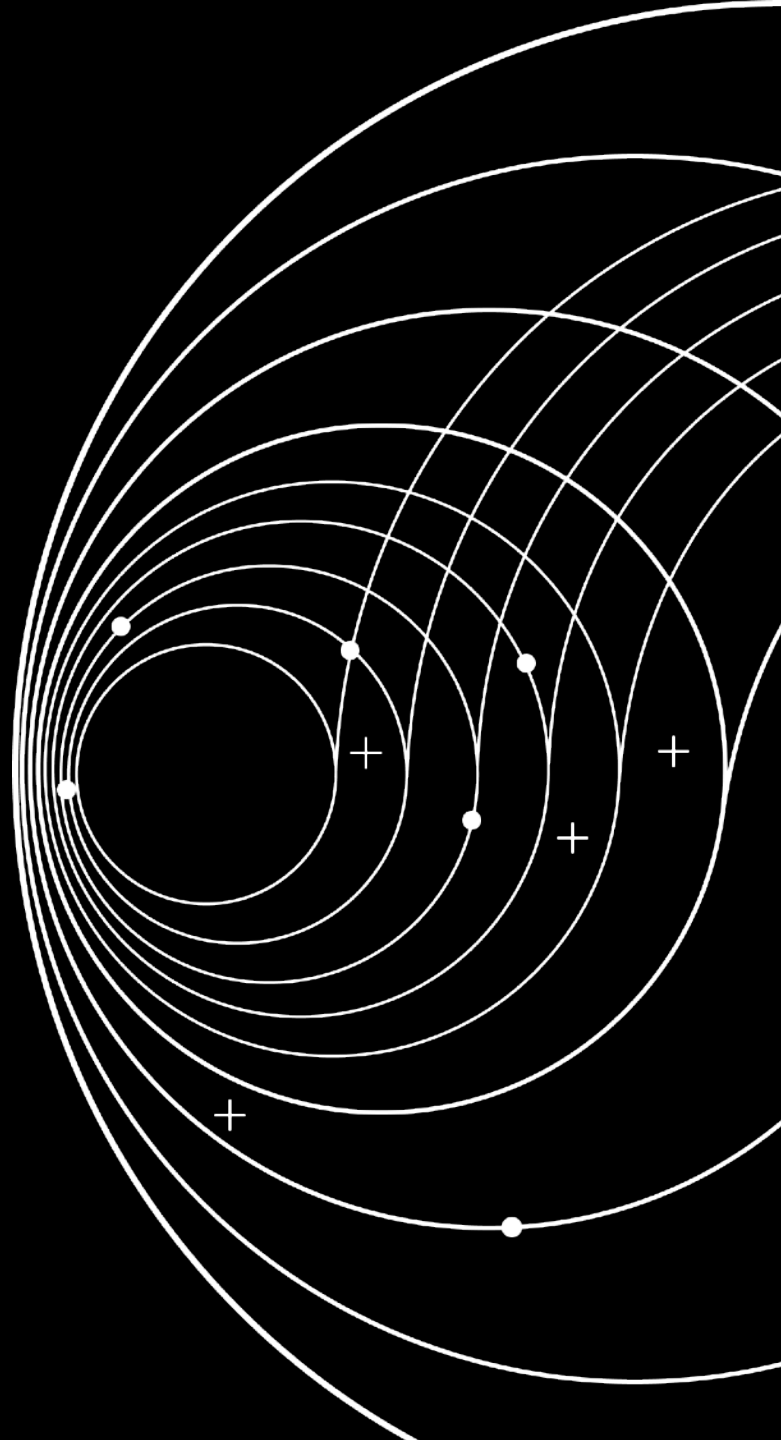
Yandex Research

Deep Learning on Tabular Data

Models, Tasks and Benchmarks



Ivan Rubachev

Tabular DL Researcher at Yandex

Hello!

Artem Babenko

Yura Gorishniy

Nikolay Kartashev

Akim Kotelnikov

Ivan Rubachev

YR Tabular DL Team

Talk Outline

01. Intro

02. DL Architectures for Tabular Data

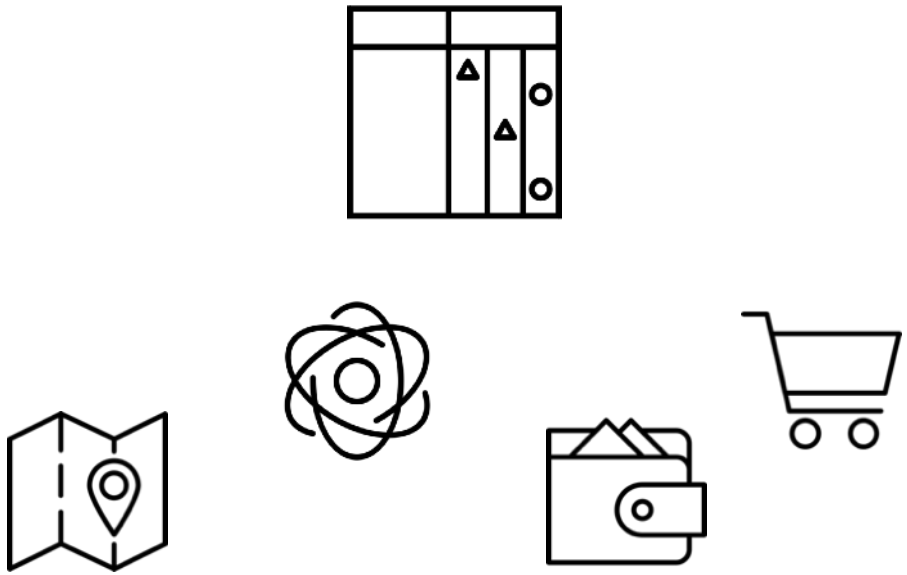
03. Tasks And Methods

04. New Benchmark

Tabular Deep Learning

Tabular Data

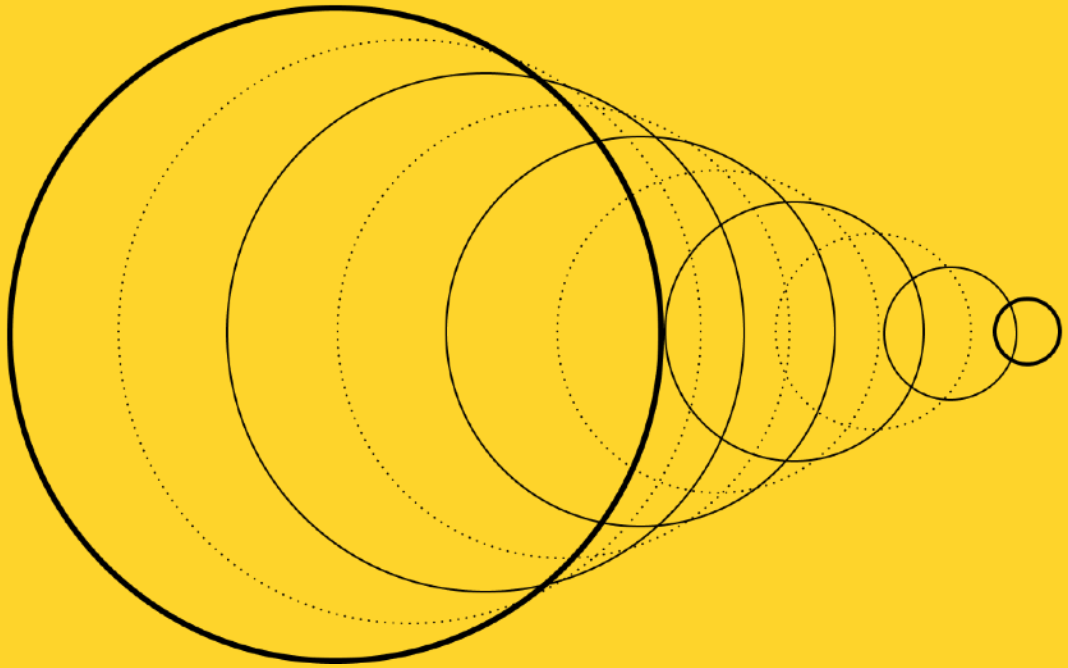
- Structured data with heterogeneous features
- Real-world applications in industry, science, medicine



Deep Learning

- Universal and popular ML toolbox
- Tackles many problems beyond vanilla supervised ML
  - Multi-table tasks
  - Multimodal neural networks
  - Generative modelling
  - Etc
- Many research questions and opportunities

Yandex Research

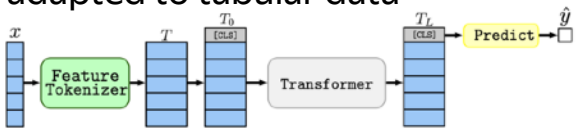


Research Recap

Architectures

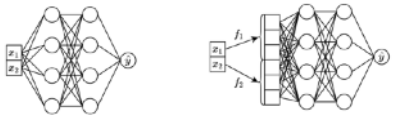
Revisiting Models [1]
FT-Transformer

- Protocols and baselines
- Transformer architecture adapted to tabular data



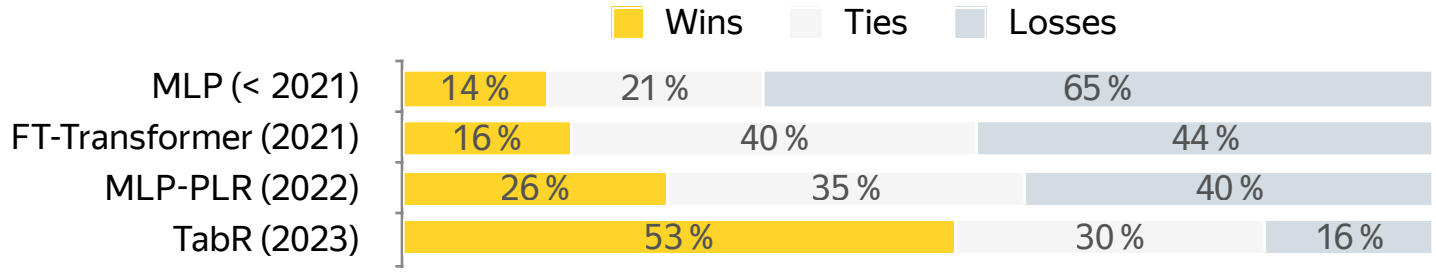
Feature Embeddings [2]
MLP-PLR

- Embedding numerical features to ease the optimization
- A universally beneficial architectural component



TabR [3]

- A retrieval-based model for tabular data
- *Strong* performance on benchmarks
- More efficient than prior approaches



Comparison to XGBoost on the academic benchmark by Grinsztajn [4]

Tasks and Methods

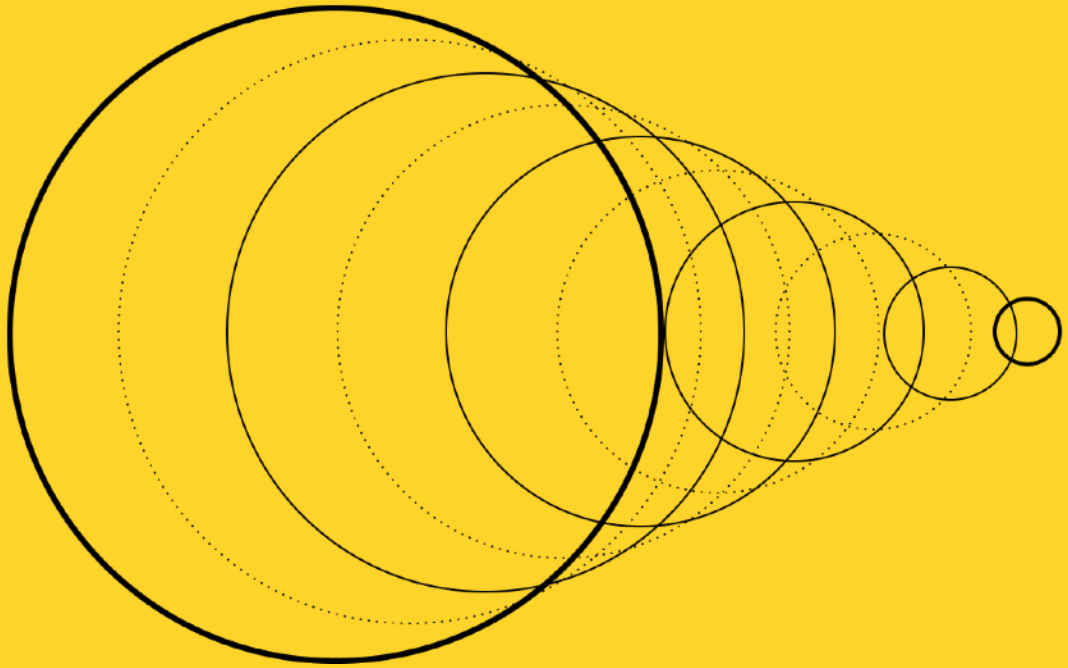
Synthetic Data Generation [5]

- Tabular data is often proprietary or private
- TabDDPM — diffusion model for tabular data generation
- Strong baseline and evaluation setup for the field

Pre-training [6]

- Simple pre-training strategies
  reconstruction, *mask-prediction*
- Trade training compute for performance
- Pre-training is beneficial on labeled data
  and smaller (10-100k) tables too

Yandex Research



New Benchmark

Let’s Look at The Academic Benchmarks

Tableshift Voting	8000	55	https://electionstudies.org/	will vote in the U.S presidential election, from a detailed questionnaire. It seems like the data goes all the way back to 1948, which makes this not realistic when not using time split		
cpu_act	8192	21	https://openml.org/d/197	This data represents logs from a server computer. The task is to predict the portion of time that cpu runs in user mode.	1	0
SpeedDating	8378	121	https://www.openml.org/search?type=data&status=active&id=40536	This dataset describes experimental speed dating events that took place from 2002 to 2004. The data describes the responses of participants to a questionnaire, and the target variable is whether they matched or not.	1	0
visualizing_soil	8641	4	https://openml.org/d/688	**Leakage.** This dataset describes a series of measurements of soil resistivity taken on a grid. The original intended target variable was the resistivity of the soil, however it wasn't the first variable, and the technical variable #1 became the target variable in the later versions of this dataset on OpenML and in the tabular benchmarks. This makes the task absurd and trivial, as a simple if between two linear transforms of two different other features in the dataset performs on par with the best algorithm mentioned in the TabR paper, beating 4 others.	0	0
yprop_4_1	8885	62	https://openml.org/d/416 and https://pubmed.ncbi.nlm.nih.gov/14502479	This dataset describes a series of chemical formulas, with a task of predicting one attribute of a molecule based on many others. The task would be better solved by graph DL methods.	0	0
Gesture Phase	9873	32	https://www.sciencedirect.com/science/article/pii/S0957417416300525	The task of this dataset is classifying gesture phases. Features are the speed and the acceleration from kinect. There are 7 videos from 3 users (3 gesture sequences from 2 and one from additional user). The paper, which introduced the dataset mentions that using same user (but different story) for evaluation influences the score. Tabular DL papers, use random split on this dataset – this is not assesing the performance on new user, not even on new sequences of one user, **not a canonical split. Without canonical split, the task contains leakage, which is easily exploited by using retrieval methods or overtuning models.**	1	0
Churn Modelling	10000	11	https://www.kaggle.com/datasets/shubh0799/churn-modelling/data	This dataset describes a set of customers of a bank, with a task of classifying whether a user will stay with the bank. Not a time split. **Unknown source** (may be synthetic). Not rich information. Narrow, No License. **No canonical split** (No time dimention)	1	0
				This dataset includes a number of simple features useful for determining	1	0

All happy ~~families~~ datasets are alike; each unhappy dataset is unhappy in its own way.

Anecdotes

SGEMM GPU kernel performance [7]

- Task is predicting the time that it takes to multiply two matrices
- Due to poor preprocessing, 3 out of 4 target variables are given with the features

Electricity [8]

- The dataset (originally named ELEC2) contains 45,312 instances dated from 7 May 1996 to 5 December 1998.
- Each example on the dataset has 5 fields, the day of week, the time stamp, the New South Wales electricity demand, the Victoria electricity demand. The scheduled electricity transfer between states and the class label
- The class label identifies the change of the price (UP or DOWN)

Academic Benchmarks

>50%

Datasets don’t handle time properly

38%

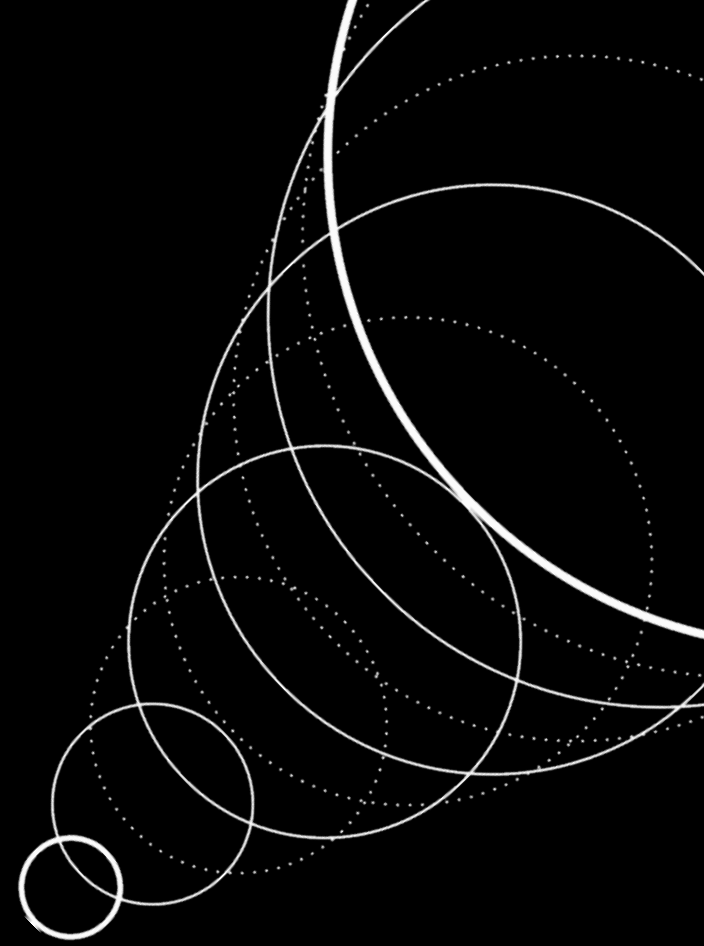
“Problematic” Datasets

~20

Features available

<1kk

Small sample sizes
Majority is bellow 100k samples



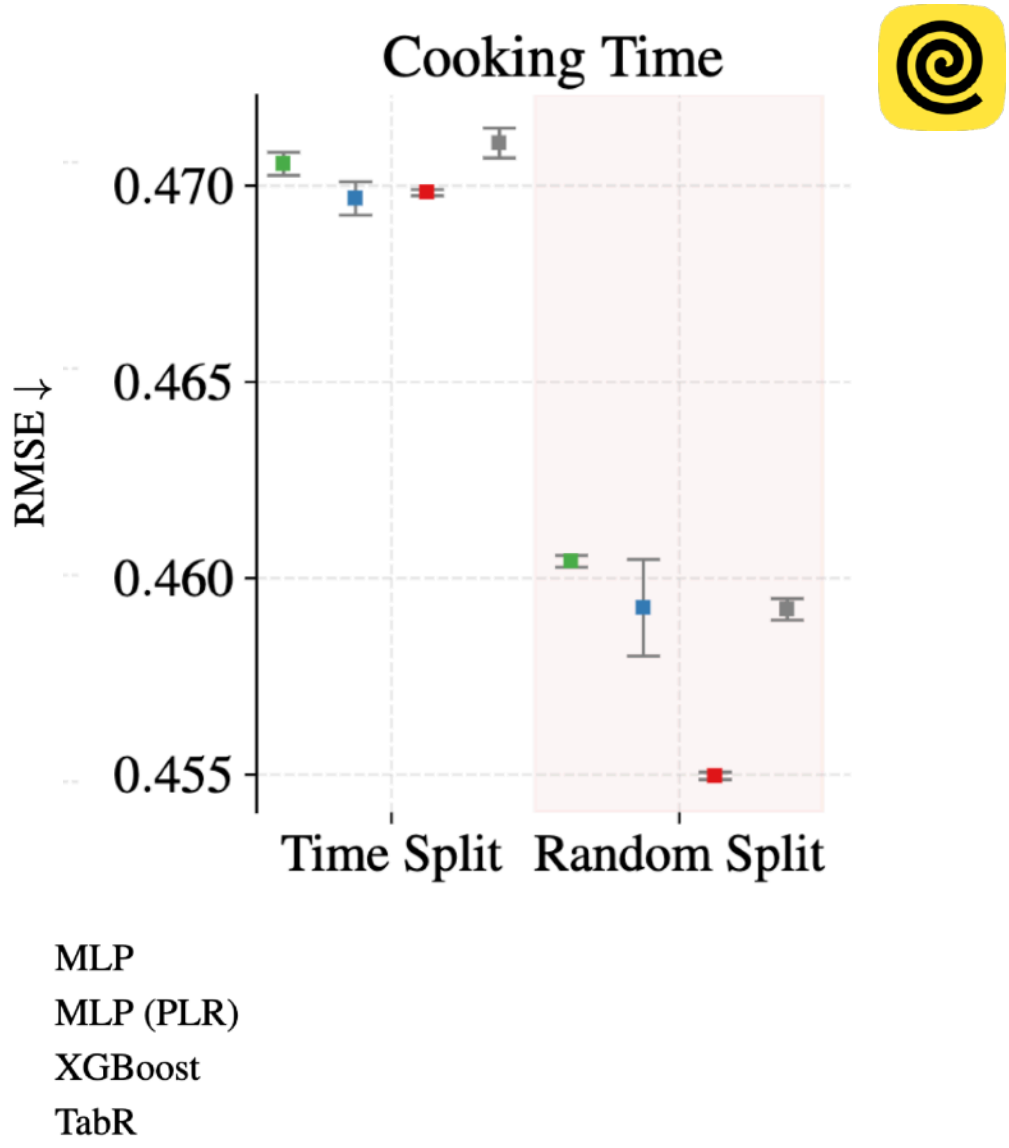
TabReD

Benchmark	Dataset Sizes (Q_{50})		Issues (#Issues / #Datasets)			Time-split		
	#Samples	#Features	Data-Leakage	Synthetic or Untraceable	Non-Tabular	Needed	Possible	Used
Grinsztajn et al. [22]	16,679	13	7 / 44	1 / 44	7 / 44	22	5	
Tabzilla [40]	3,087	23	3 / 36	6 / 36	12 / 36	12	0	
WildTab [35]	546,543	10	1* / 3	1 / 3	0 / 3	1	1	✗
TableShift [18]	840,582	23	0 / 15	0 / 15	0 / 15	15	8	
Gorishniy et al. [21]	57,909	20	1* / 10	1 / 10	0 / 10	7	1	
TabReD (ours)	7,163,150	261	✗	✗	✗	✓	✓	✓

No benchmark beside TabReD focuses on temporal-shift based evaluation, less.
OpenML based datasets have more quality issues

Temporal shift

- GBDT’s are less robust to temporal shift
- Realistic evaluation setups are important for healthy progress

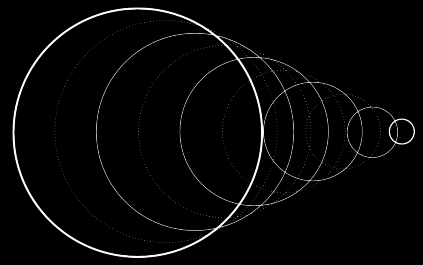


Summary

- A new benchmark with datasets, closer resembling real-world scenarios
- Sources: Kaggle and Yandex Eats, Maps, Weather, Lavka
- Datasets with 10M samples and feature-engineering *(with up-to 1000s of features)*
- All datasets have timestamps

Experimental results

- Performance differences are less pronounced (feature-engineering)
- Time-splits are important
- General Progress Transfers
- MLP-PLR and GBDTs - top-2 models



Thanks! Any Questions? References:

- [1] Gorishniy, Yury, et al. "Revisiting deep learning models for tabular data." *Advances in Neural Information Processing Systems* 34 (2021): 18932-18943.
- [2] Gorishniy, Yury, Ivan Rubachev, and Artem Babenko. "On embeddings for numerical features in tabular deep learning." *Advances in Neural Information Processing Systems* 35 (2022): 24991-25004.
- [3] Gorishniy, Yury, et al. "Tabr: Unlocking the power of retrieval-augmented tabular deep learning." *arXiv preprint arXiv:2307.14338* (2023).
- [4] Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?." *Advances in neural information processing systems* 35 (2022): 507-520.
- [5] Kotelnikov, Akim, et al. "Tabddpm: Modelling tabular data with diffusion models." *International Conference on Machine Learning*. PMLR, 2023.
- [6] Rubachev, Ivan, et al. "Revisiting pretraining objectives for tabular deep learning." *arXiv preprint arXiv:2207.03208* (2022).
- [7] SGEMM GPU kernel performance https://www.openml.org/d/42963
- [8] Electricity https://www.openml.org/d/151