

NeurIPS 2024 – Datasets and Benchmarks

TabularBench: Benchmarking Adversarial Robustness for Tabular Deep Learning in Real-world Use-cases

Thibault Simonetto¹, Salah Ghamizi^{2, 3}, Maxime Cordy¹

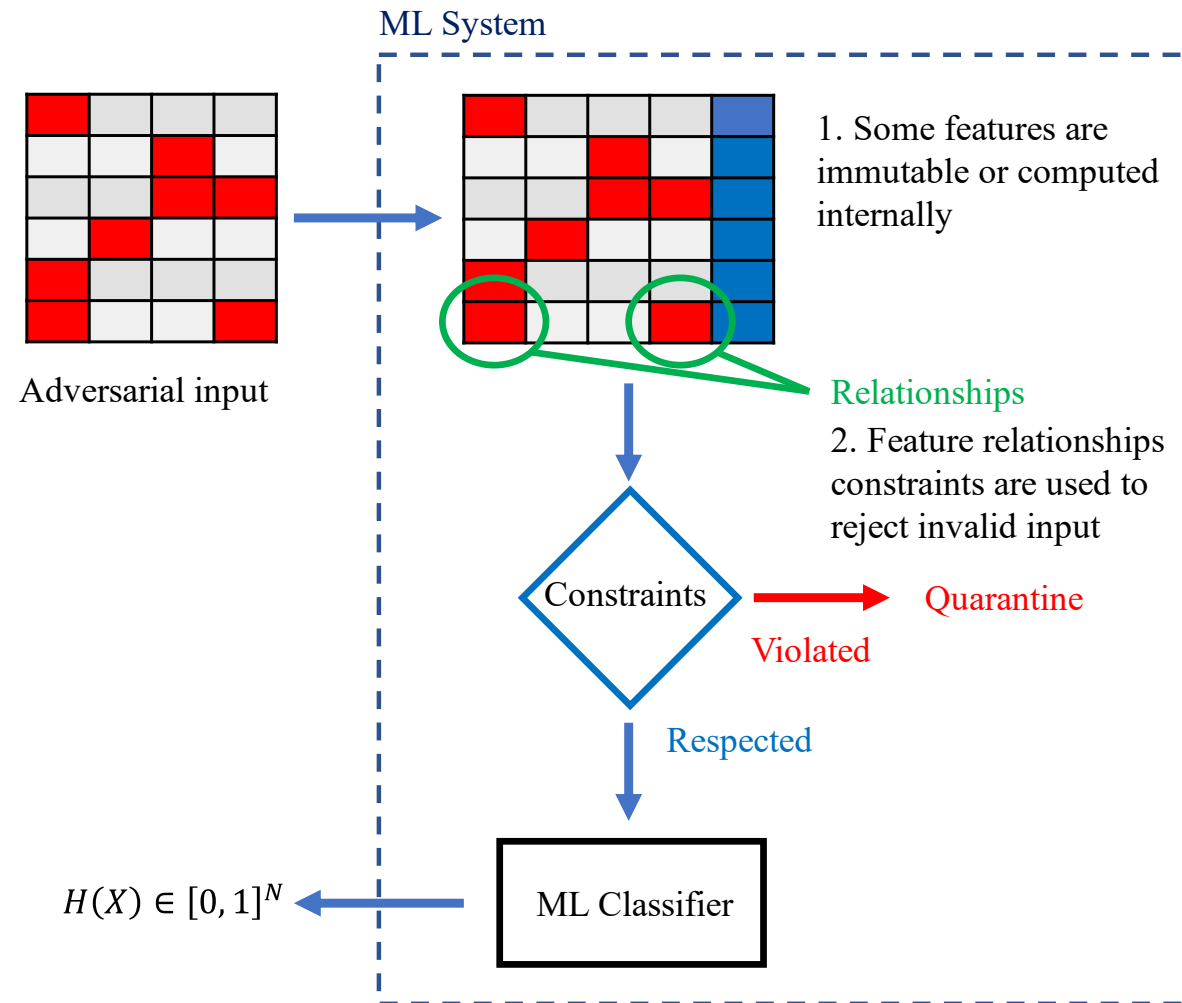
¹University of Luxembourg, Luxembourg

²Luxembourg Institute of Science and Technology, Luxembourg

³RIKEN Center for Advanced Intelligence Project, Tokyo, Japan



Adversarial examples in tabular data



Relation constraints on feature space

Finance:

avg transaction amount $\leq$ max transaction amount

$$\textit{installment} = \textit{loan_amount} \times \frac{\textit{int_rate} \times (1 + \textit{int_rate})^{\textit{term}}}{(1 + \textit{int_rate})^{\textit{term}} - 1}$$

Constrained adversarial attacks

Given

- a classification model H ,
- a maximum perturbation ϵ under a L_p distance
- a set of **constraints** Ω

Objective of **constrained** adversarial attacks, for clean sample x find perturbation δ :

✓ With $H(x) \neq H(x + \delta)$

✓ With $L_p(x, x + \delta) < \epsilon$

✓ $x + \delta \models \Omega$

Experimental settings

Datasets:



Credit scoring
Lending Club Loan Data



Botnet detection
CTU



URL phishing
URL



Malware detection
MALWARE



ICU survival
WIDS

Models: 5 Neural network architectures

- 2 Regularizations
- 2 Transformers
- 1 Semi-supervised

Attack: Constrained Adaptive Attack [1]:

CAPGD



MOEVA

[1] Thibault Simonetto, Salah Ghamizi, and Maxime Cordy. "Constrained Adaptive Attack: Effective Adversarial Attack Against Deep Neural Networks for Tabular Data" in Advances in Neural Information Processing Systems, 2024.

Madry Adversarial training

Saddle point problem

$$\arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \mathcal{S}} l(\theta, x + \delta, y) \right]$$

Training

Adversarial

Adversarial part solved using PGD

$$x^{(k+1)} = P_{\mathcal{S}}(x^{(k)} + \eta^{(k)} \nabla l(h(x), y))$$

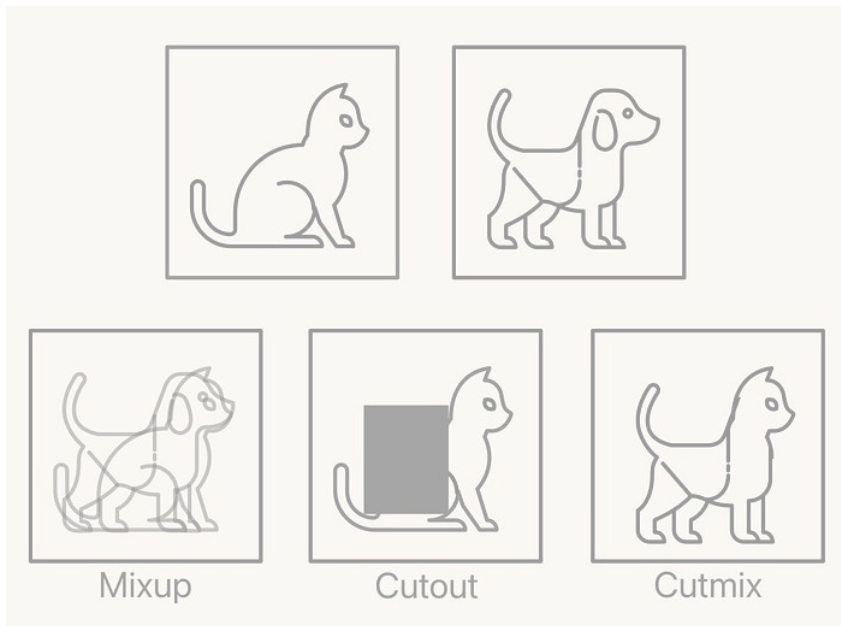
Regularization via mix of clean and adversarial examples

$$\hat{\mathcal{L}}_B = \frac{1}{(m-k) + \lambda k} \left(\lambda \sum_{i=1}^k \mathcal{L}(h(PGD(x_i, y_i)), y_i) + \sum_{i=k+1}^m \mathcal{L}(h(x_i), y_i) \right)$$

Data augmentation with Adversarial training

$$\mathcal{L}_B = \frac{1}{\sum_{j=1}^4 \lambda_j k_j} \left(\begin{aligned} &\lambda_1 \sum_{i=1}^{k_1} \mathcal{L}(h(x_i, y_i), y_i) + && \text{Clean examples} \\ &\lambda_2 \sum_{i=k_1+1}^{k_1+k_2} \mathcal{L}(h(PGD(x_i, y_i)), y_i) + && \text{Adversarial examples} \\ &\lambda_3 \sum_{i=k_1+k_2+1}^{k_1+k_2+k_3} \mathcal{L}(h(GEN(x_i, y_i)), y_i) + && \text{Synthetic examples} \\ &\lambda_4 \sum_{i=k_1+k_2+k_3+1}^m \mathcal{L}(h(PGD(GEN(x_i, y_i)), y_i), y_i) \end{aligned} \right) && \text{Synthetic and Adversarial examples}$$

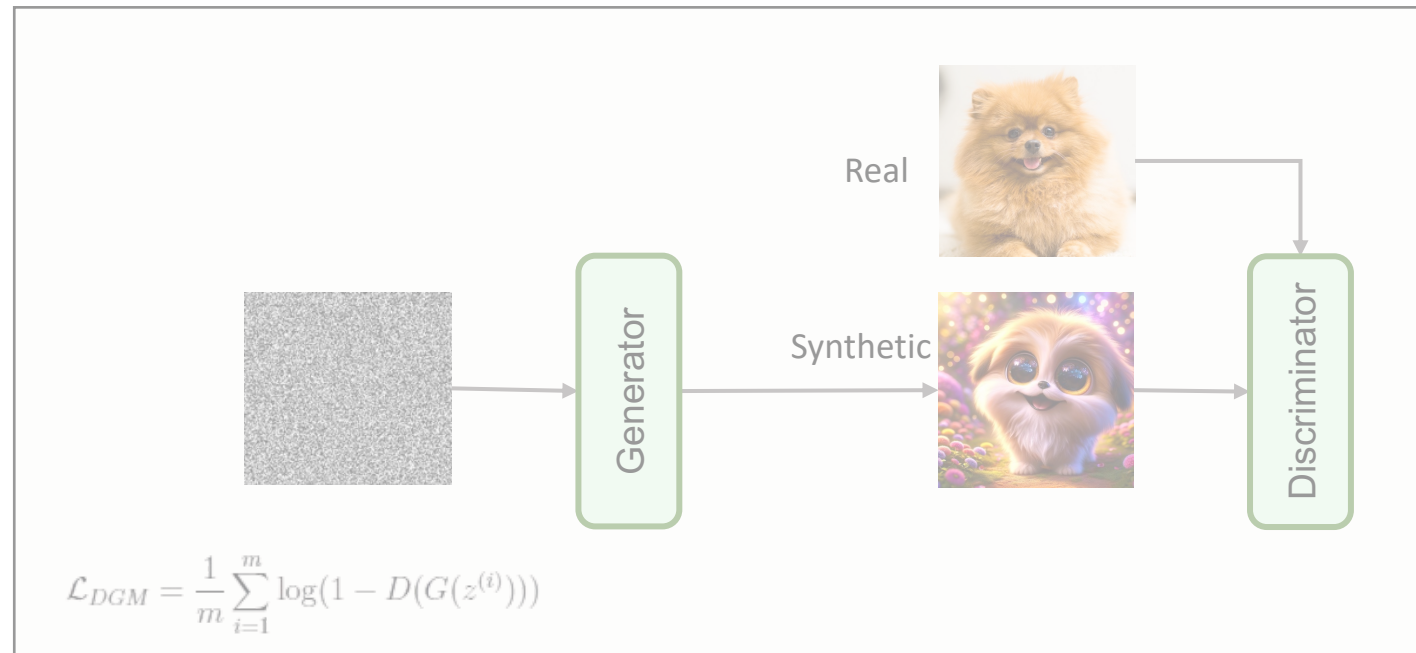
Data augmentations



<https://towardsdatascience.com/cutout-mixup-and-cutmix-implementing-modern-image-augmentations-in-pytorch-a9d7db3074ad>

Heuristic-based

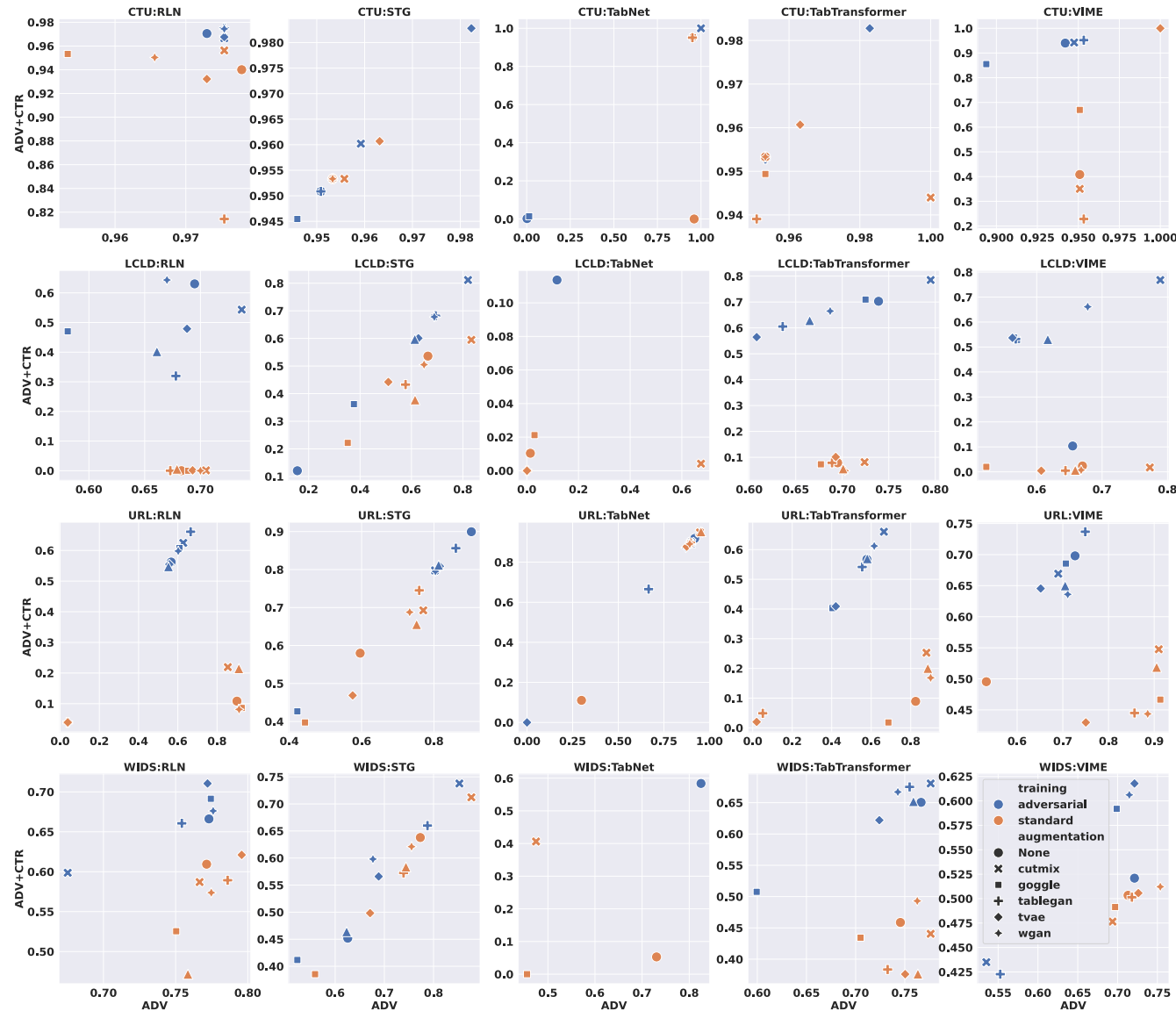
- Cutmix



Deep Generative Models

- CTGAN, TVAE, GOGGLE, TableGAN, WGAN

Large scale empirical study



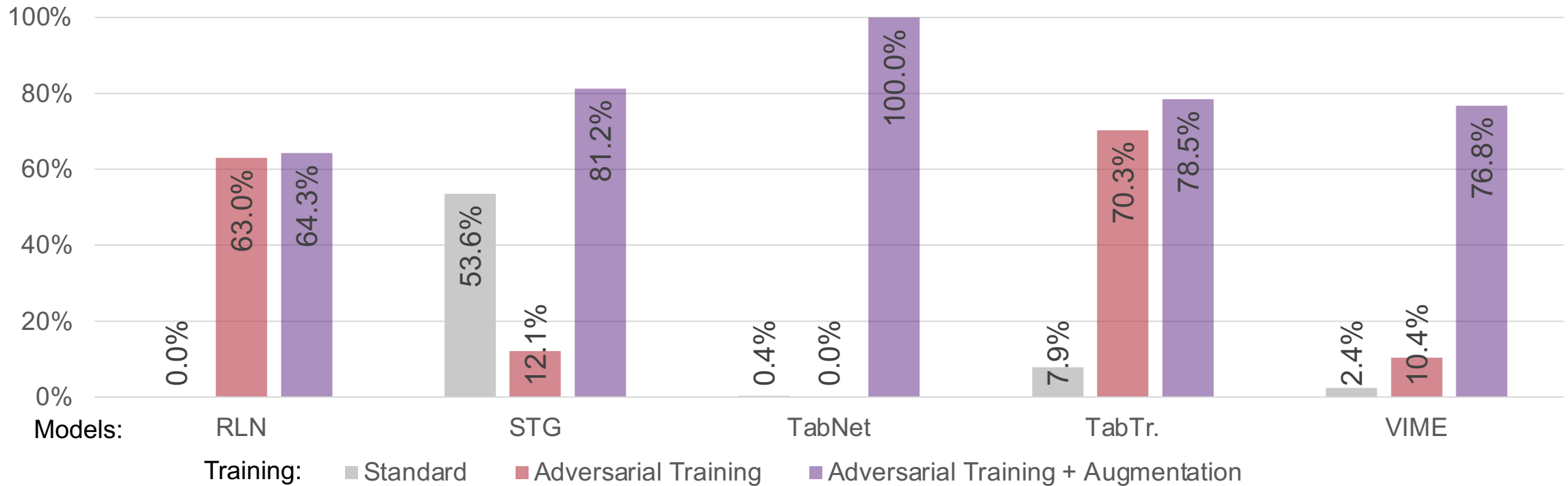
Clean accuracy

Robust accuracy

- Constrained
- Unconstrained

Impact on robust accuracy

Robust accuracy



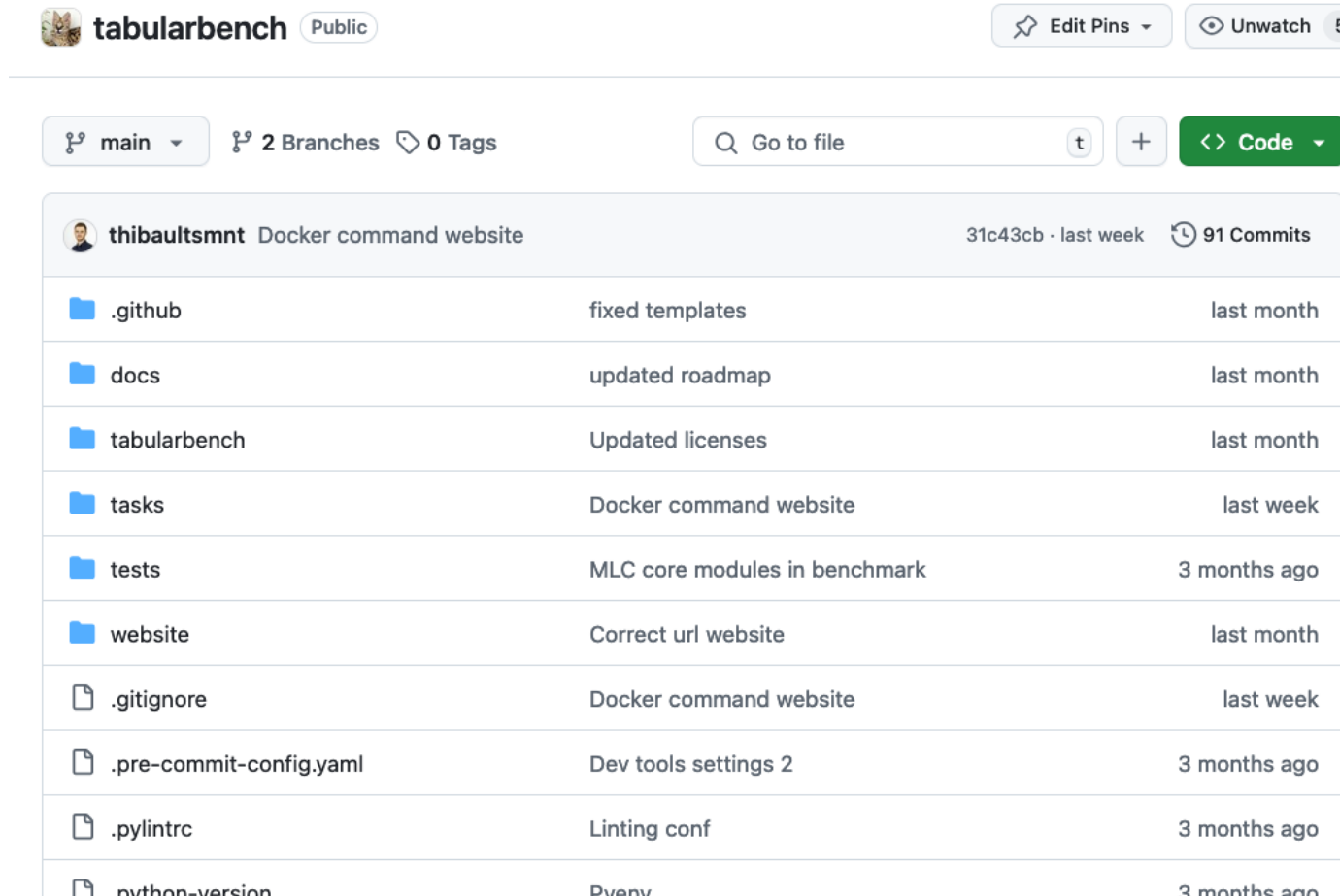
Dataset: Lending Club Loan data

Insight:

Adversarial training with data augmentation outperforms adversarial training alone.

TabularBench: A benchmark for Tabular robustness

Source code of models, attacks, dataset, constraints



thibaultsmnt Public Edit Pins Unwatch 5

main 2 Branches 0 Tags Go to file t + <> Code

thibaultsmnt	Docker command website	31c43cb · last week	91 Commits
.github	fixed templates	last month	
docs	updated roadmap	last month	
tabularbench	Updated licenses	last month	
tasks	Docker command website	last week	
tests	MLC core modules in benchmark	3 months ago	
website	Correct url website	last month	
.gitignore	Docker command website	last week	
.pre-commit-config.yaml	Dev tools settings 2	3 months ago	
.pylintrc	Linting conf	3 months ago	
python-version	Python	3 months ago	

<https://github.com/serval-uni-lu/tabularbench>

TabularBench: A benchmark for Tabular robustness

Leaderboard: evaluation of 200+ models

[tabularbench](#)

TabularBench: Adversarial robustness benchmark for tabular data [Documentation](#)

Leaderboard

You are currently viewing results of the leaderboard. View MOEVA attack results [here](#).

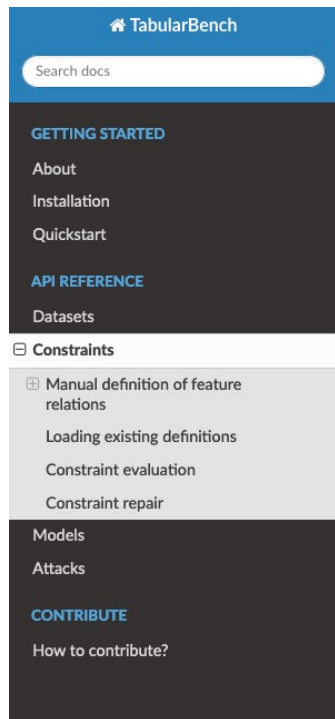
CTU [↗](#)

Search:								
architecture ↕	training ↕	augmentation ↕	ID ↕	ADV+CTR ↕	ADV ↕	auc ↕	accuracy ▲	precision
TabNet	adversarial	tvae	1	1	1	0.976447	0.00738898	0.00738898
VIME	adversarial	ctgan	1	1	1	0.741099	0.00738898	0.00738898
VIME	adversarial	tvae	1	1	1	0.727257	0.00738898	0.00738898
VIME	standard	ctgan	1	1	1	0.972449	0.00738898	0.00738898
VIME	standard	tvae	1	1	1	0.949582	0.00751607	0.007389
TabNet	adversarial	ctgan	1	1	1	0.976819	0.0156676	0.007450
TabTransformer	standard	ctgan	1	0.94398	1	0.630416	0.0437893	0.00766
TabTransformer	adversarial	ctgan	1	0.94398	1	0.627151	0.0448604	0.007676
STG	adversarial	tvae	0.982801	0.982801	0.98231	0.981094	0.435641	0.01270
TabTransformer	adversarial	tvae	0.982801	0.982801	0.982801	0.974144	0.608674	0.01822
STG	standard	tvae	0.963145	0.960688	0.963145	0.984115	0.890109	0.06096
STG	adversarial	ctgan	0.960688	0.960688	0.960688	0.986310	0.920578	0.00101

<https://serval-uni-lu.github.io/tabularbench/>

TabularBench: A benchmark for Tabular robustness

Simplified contributions



Manual definition of feature relations

all classes below are defined in `tabularbench.constraints.relation_constraint`.

Constraints between features can be expressed in natural language. For example, we express the constraint $F_0 = F_1 + F_2$ such as:

```
from tabularbench.constraints.relation_constraint import Feature
constraint1 = Feature(0) == Feature(1) + Feature(2)
```

<https://serval-uni-lu.github.io/tabularbench/doc/constraints.html>

TabularBench: A benchmark for Tabular robustness

Simplified contributions



Add a title

[New model]: |

Model name

Provide a name for the model.

e.g., TabTransformer Cutmix Model

Paper & authors references

Provide a name of the associated paper with the relevant bibtex

e.g., @article{simonetto2024constrained, title={Constrained Adaptive Attack: Effective Adversarial Attack Against Deep Neural Networks for Tabular Data}, author={Simonetto, Thibault and Ghamizi, Salah and Cordy, Maxime}, journal={arXiv preprint arXiv:2406.00775}, year={2024} }

Leaderboard claim(s)

Add here the claim for your model.

* Architecture:
* Dataset:
* Eps:
* Clean accuracy:

Share my model in the model zoo

<https://github.com/serval-uni-lu/tabularbench/issues>

TabularBench: A benchmark for Tabular robustness

Towards building robust machine learning models for Constrained Tabular Data

Checkout our
benchmark



5 constrained datasets
- Up to 360 constraints

5 model architectures

14 training methods

290 pretrained models

Available on pip, conda and docker