RealMLP: Advancing MLPs and default parameters for tabular data

David Holzmüller[1] [0000−0002−9443−0049], Léo Grinsztajn[2] [0000−0002−4436−7109], and Ingo Steinwart[3] [0000−0002−4006−8435]

[1] INRIA, Ecole Normale Superieure, PSL University
[2] INRIA
[3] University of Stuttgart

Abstract. For classification and regression on tabular data, the dominance of gradient-boosted decision trees (GBDTs) has recently been challenged by often much slower deep learning methods with extensive hyperparameter tuning. We address this discrepancy by introducing (a) RealMLP, an improved multilayer perceptron (MLP), and (b) strong meta-tuned default parameters for GBDTs and RealMLP. We tune RealMLP and the default parameters on a meta-train benchmark with 118 datasets and compare them to hyperparameter-optimized versions on a disjoint meta-test benchmark with 90 datasets, as well as the GBDT-friendly benchmark by Grinsztajn et al. (2022). Our benchmark results on medium-to-large tabular datasets (1K–500K samples) show that RealMLP offers a favorable time-accuracy tradeoff compared to other neural baselines and is competitive with GBDTs in terms of benchmark scores. Moreover, a combination of RealMLP and GBDTs with improved default parameters can achieve excellent results without hyperparameter tuning. Finally, we demonstrate that some of RealMLP's improvements can also considerably improve the performance of TabR with default parameters.

Keywords: Tabular data · Neural networks · Default parameters.

1 Overview

Tabular data is a wide-spread data modality in practice, yet the rapid progress of deep learning on images and text data has struggled to translate to tabular data, where gradient-boosted trees have dominated benchmarks. Compared to gradient boosting, deep learning models for tabular data offer advantages in flexibility, allowing to more easily incorporate multi-modal and multi-table data, incorporate pre-training and semi-supervised learning, compute derivatives, handle high-dimensional output spaces, and more. However, many state-of-the-art deep learning methods for tabular data are slow to train, especially when combined with hyperparameter tuning, and yet struggle to beat gradient-boosted decision trees in benchmarks [5, 7, 8].

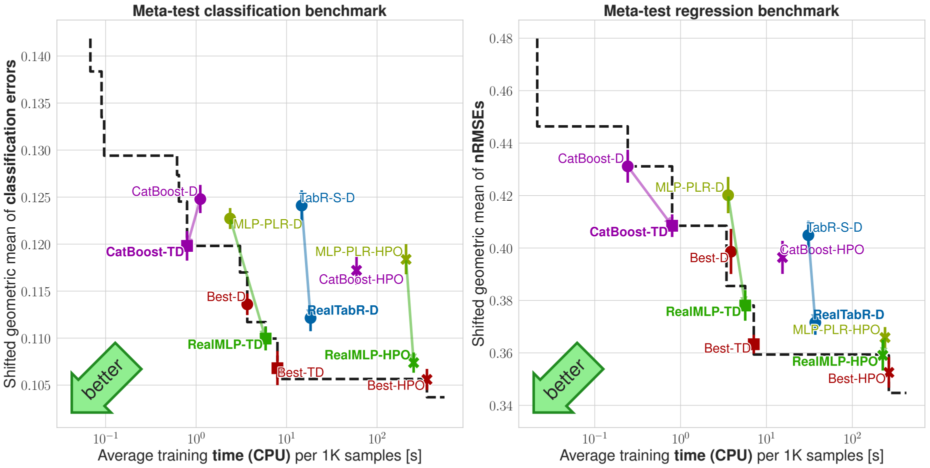

Fig. 1. Benchmark results of selected methods on the meta-test benchmark. The y-axis shows geometric_mean(err + 0.01), where err is the classification error or normalized RMSE. The x-axis shows average runtimes per 1K samples on a CPU.

To alleviate this issue, in [6], we introduce *RealMLP*, a multilayer perceptron (MLP) with improvements in architecture, preprocessing, training, regularization, and initialization. RealMLP achieves excellent results on multiple benchmarks and *takes the first place in overall rank on the independent 300-dataset benchmark by Ye et al. [8]*. Moreover, we show that RealMLP already achieves strong results with its default parameters, which we optimized on a meta-train benchmark of 118 datasets and evaluated on a disjoint benchmark of 90 datasets selected using well-defined criteria from the AutoML benchmark [2] and the CTR23 benchmark [1].

We provide multiple ablation studies to investigate the benefit of different improvements in RealMLP, showing that architecture is important but other choices matter a lot as well. In addition, we show that many of our improvements to RealMLP are also beneficial for the recent retrieval-based TabR model [4], resulting in the RealTabR model. In addition, we study the benefits of performing algorithm selection over RealMLP and boosted trees, leading to the "Best" models in Figure 1, as well as meta-learned defaults for boosted trees. Figure 1 shows results on the meta-test benchmark for different methods with library defaults (D), (meta-learned) tuned defaults (TD), and hyperparameter optimization (HPO). RealMLP and RealTabR achieve excellent results, outperforming CatBoost-HPO as well as their respective baselines MLP-PLR [3] and TabR. We provide scikit-learn interfaces for all investigated methods at

github.com/dholzmueller/pytabkit

Bibliography

[1] Fischer, S.F., Feurer, M., Bischl, B.: OpenML-CTR23–A curated tabular regression benchmarking suite. In: AutoML Conference 2023 (Workshop) (2023)

[2] Gijsbers, P., Bueno, M.L., Coors, S., LeDell, E., Poirier, S., Thomas, J., Bischl, B., Vanschoren, J.: AMLB: an AutoML benchmark. Journal of Machine Learning Research **25**(101), 1–65 (2024), https://www.jmlr.org/papers/v25/22-0493.html

[3] Gorishniy, Y., Rubachev, I., Babenko, A.: On embeddings for numerical features in tabular deep learning. Neural Information Processing Systems (2022)

[4] Gorishniy, Y., Rubachev, I., Kartashev, N., Shlenskii, D., Kotelnikov, A., Babenko, A.: TabR: Tabular deep learning meets nearest neighbors. In: International Conference on Learning Representations (2024)

[5] Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? Neural Information Processing Systems (2022)

[6] Holzmüller, D., Grinsztajn, L., Steinwart, I.: Better by default: Strong pretuned MLPs and boosted trees on tabular data. In: Neural Information Processing Systems (2024)

[7] McElfresh, D., Khandagale, S., Valverde, J., Prasad C, V., Ramakrishnan, G., Goldblum, M., White, C.: When do neural nets outperform boosted trees on tabular data? In: Neural Information Processing Systems (2023)

[8] Ye, H.J., Liu, S.Y., Cai, H.R., Zhou, Q.L., Zhan, D.C.: A closer look at deep learning on tabular data. arXiv:2407.00956 (2024)