Advancing tabular machine learning models and benchmarks

SODA Kick-off

David Holzmüller

June 3rd, 2025

Tabular data: Motivation

Age	Blood pressure	Smoker?	Disease	Readmitted?
45	120	No	Cancer	No
60	140	Yes	Asthma	Yes
38	130	N/A	AIDS	No
N/A	160	Yes	Stroke	**?**

Tabular data problems are wide-spread in science & industry

State of neural networks in 2022

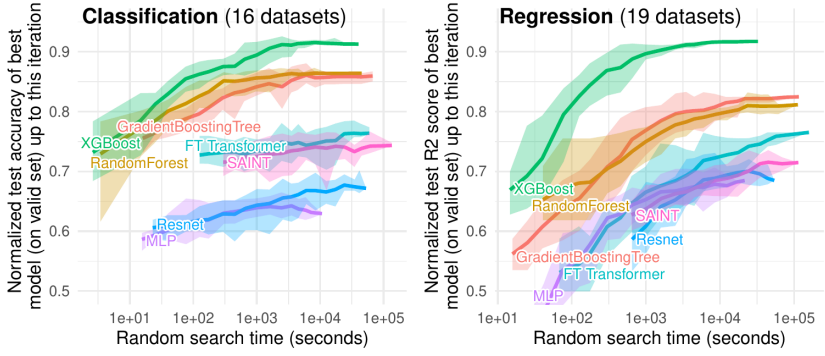
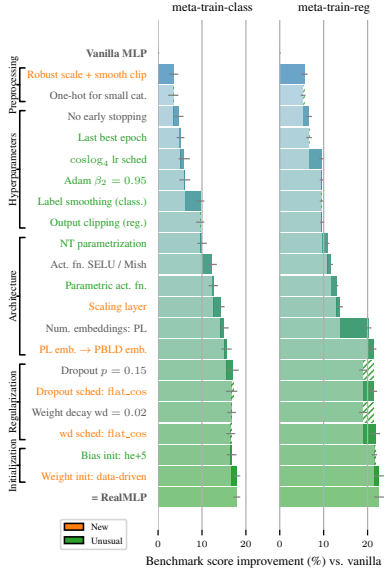


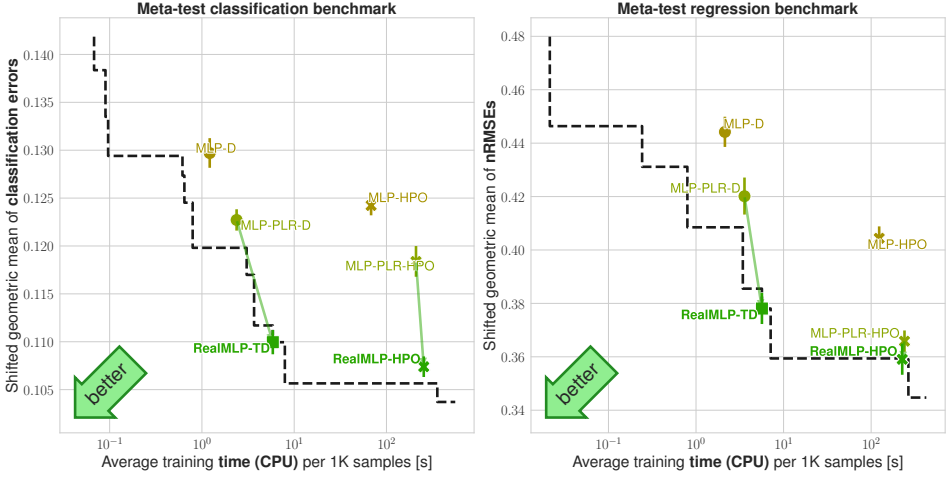
Figure 6: **Time benchmark on medium-sized datasets, with only numerical features**. The first random search iteration corresponds to default hyperparameters. Each value corresponds to the test score of the best model (on the validation set) after a specific time spent doing random search, averaged on 15 shuffles of the random search order. The ribbon corresponds to the minimum and maximum scores on these 15 shuffles.

Source: Grinsztajn et al. (2022)

RealMLP: a modern MLP (Holzmüller et al., 2024)

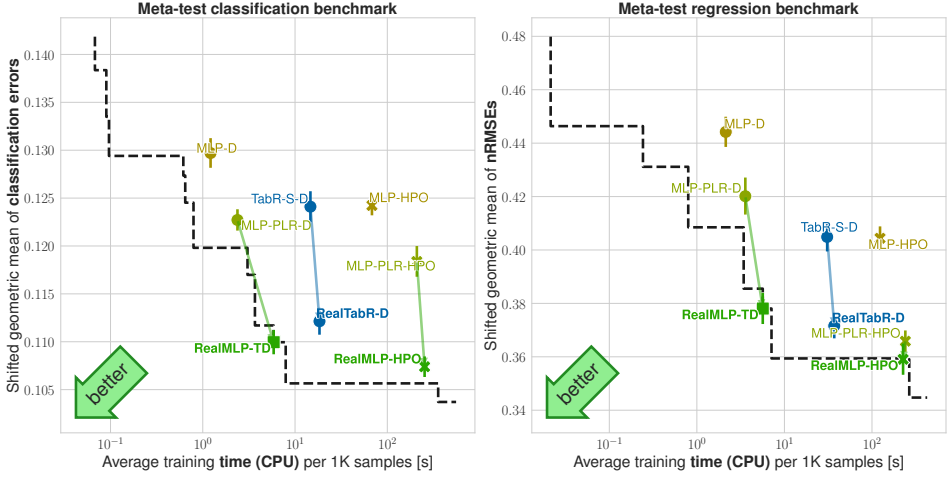


Results on separate benchmark (Holzmüller et al., 2024)



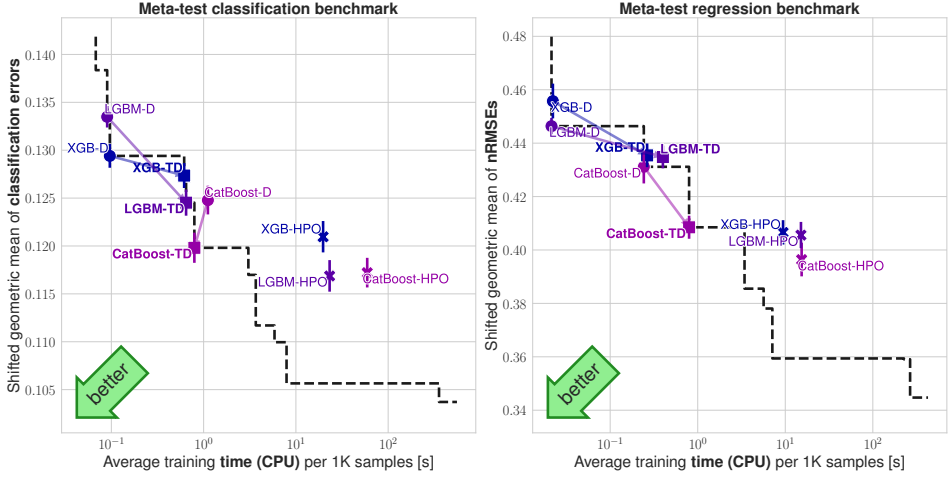
Library: `github.com/dholzmueller/pytabkit`

Results on separate benchmark (Holzmüller et al., 2024)



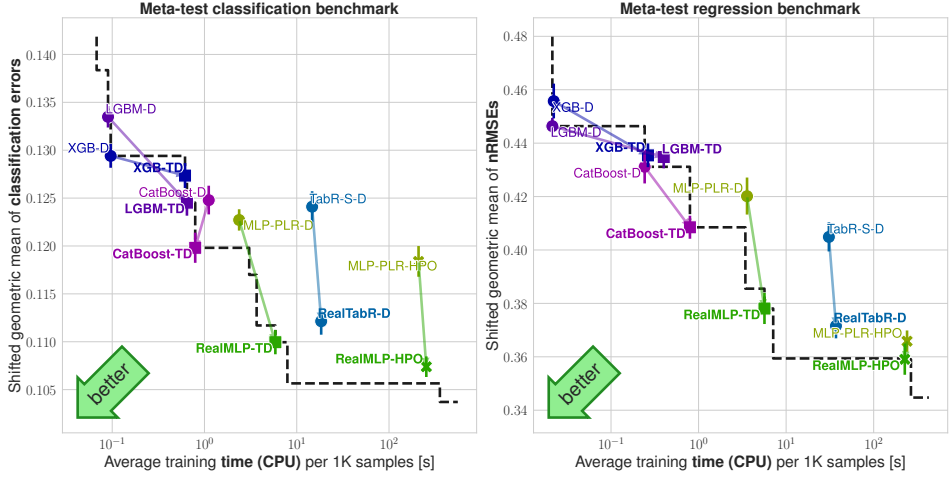
Library: `github.com/dholzmueller/pytabkit`

Results on separate benchmark (Holzmüller et al., 2024)



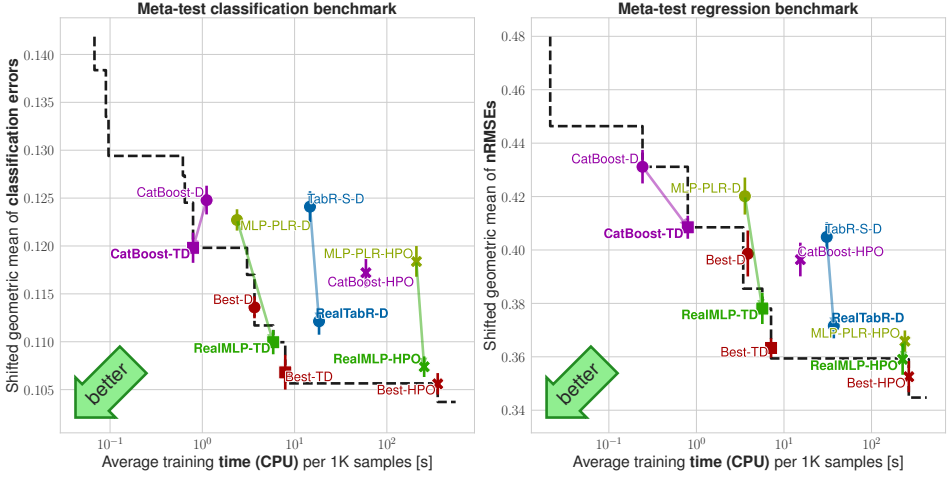
Library: `github.com/dholzmueller/pytabkit`

Results on separate benchmark (Holzmüller et al., 2024)



Library: `github.com/dholzmueller/pytabkit`

Results on separate benchmark (Holzmüller et al., 2024)



Library: `github.com/dholzmueller/pytabkit`

TabArena: A living benchmark for tabular ML

Upcoming paper with Nick Erickson, Lennart Purucker, Andrej Tschalzev, me, Prateek Desai, David Salinas, Frank Hutter

TabArena: A living benchmark for tabular ML

Upcoming paper with Nick Erickson, Lennart Purucker, Andrej Tschalzev, me, Prateek Desai, David Salinas, Frank Hutter

- Living benchmark: online leaderboard (`tabarena.ai`), continuous maintenance

TabArena: A living benchmark for tabular ML

Upcoming paper with Nick Erickson, Lennart Purucker, Andrej Tschalzev, me, Prateek Desai, David Salinas, Frank Hutter

- Living benchmark: online leaderboard (`tabarena.ai`), continuous maintenance
- Strict selection for tabular predictive IID datasets (51 datasets)

TabArena: A living benchmark for tabular ML

Upcoming paper with Nick Erickson, Lennart Purucker, Andrej Tschalzev, me, Prateek Desai, David Salinas, Frank Hutter

- Living benchmark: online leaderboard (`tabarena.ai`), continuous maintenance
- Strict selection for tabular predictive IID datasets (51 datasets)
- Best practices for model evaluation (inner and outer cross-validation)

TabArena: A living benchmark for tabular ML

Upcoming paper with Nick Erickson, Lennart Purucker, Andrej Tschalzev, me, Prateek Desai, David Salinas, Frank Hutter

- Living benchmark: online leaderboard (`tabarena.ai`), continuous maintenance
- Strict selection for tabular predictive IID datasets (51 datasets)
- Best practices for model evaluation (inner and outer cross-validation)
- Strong baselines (optimized search spaces for tree-based models)

TabArena: A living benchmark for tabular ML

Upcoming paper with Nick Erickson, Lennart Purucker, Andrej Tschalzev, me, Prateek Desai, David Salinas, Frank Hutter

- Living benchmark: online leaderboard (`tabarena.ai`), continuous maintenance
- Strict selection for tabular predictive IID datasets (51 datasets)
- Best practices for model evaluation (inner and outer cross-validation)
- Strong baselines (optimized search spaces for tree-based models)
- More holistic classification metrics (AUC / log-loss)

TabArena: A living benchmark for tabular ML

Upcoming paper with Nick Erickson, Lennart Purucker, Andrej Tschalzev, me, Prateek Desai, David Salinas, Frank Hutter

- Living benchmark: online leaderboard (`tabarena.ai`), continuous maintenance
- Strict selection for tabular predictive IID datasets (51 datasets)
- Best practices for model evaluation (inner and outer cross-validation)
- Strong baselines (optimized search spaces for tree-based models)
- More holistic classification metrics (AUC / log-loss)
- Evaluation also with weighted ensembling (Caruana et al., 2004)

TabArena: A living benchmark for tabular ML

Upcoming paper with Nick Erickson, Lennart Purucker, Andrej Tschalzev, me, Prateek Desai, David Salinas, Frank Hutter

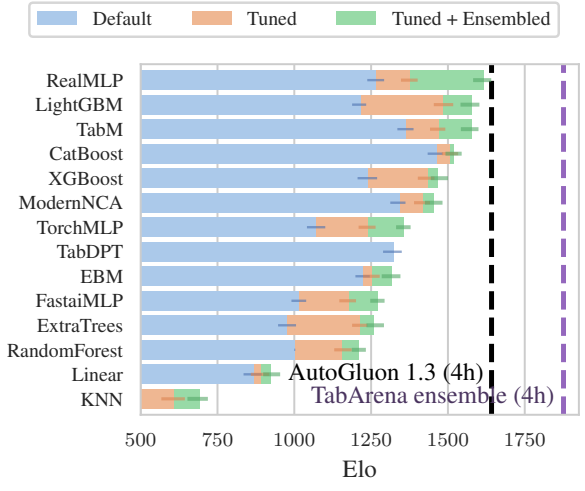
- Living benchmark: online leaderboard (`tabarena.ai`), continuous maintenance
- Strict selection for tabular predictive IID datasets (51 datasets)
- Best practices for model evaluation (inner and outer cross-validation)
- Strong baselines (optimized search spaces for tree-based models)
- More holistic classification metrics (AUC / log-loss)
- Evaluation also with weighted ensembling (Caruana et al., 2004)
- Storing the predictions for weighted ensembling

TabArena: A living benchmark for tabular ML

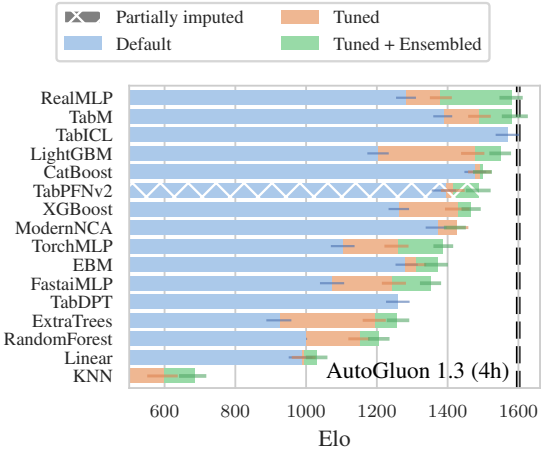
Upcoming paper with Nick Erickson, Lennart Purucker, Andrej Tschalzev, me, Prateek Desai, David Salinas, Frank Hutter

- Living benchmark: online leaderboard (`tabarena.ai`), continuous maintenance
- Strict selection for tabular predictive IID datasets (51 datasets)
- Best practices for model evaluation (inner and outer cross-validation)
- Strong baselines (optimized search spaces for tree-based models)
- More holistic classification metrics (AUC / log-loss)
- Evaluation also with weighted ensembling (Caruana et al., 2004)
- Storing the predictions for weighted ensembling
- Main aggregation: Elo (a rating system for predicting win-rates vs other algorithms)

TabArena: Main results



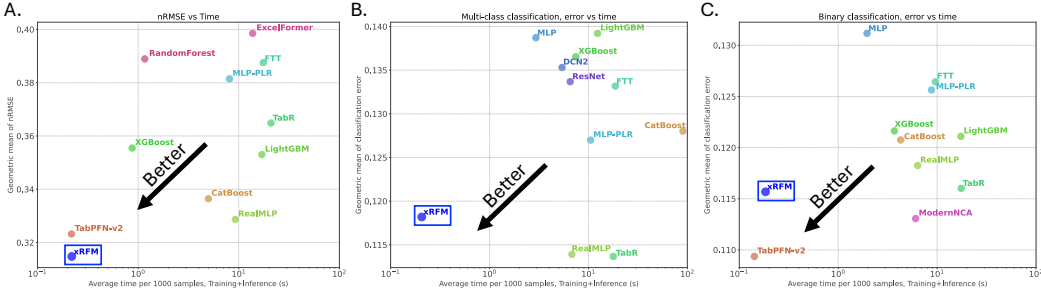
TabICL: A foundation model for large data (Qu et al., 2025)



Results for TabICL-compatible datasets

Coming soon: Strong kernel methods for tabular data

Upcoming paper with Daniel Beaglehole, Adityanarayanan Ramakrishnan, Mikhail Belkin



xRFM: improved recursive feature machines (Radhakrishnan et al., 2024)

Plans for extending tabular foundation models

- Direct extensions: Regression, missing values, etc.
- More diverse inputs (text, multi-table, prior knowledge)
- More diverse outputs (full posteriors, causal estimates?)

Conclusion

- More model types for the tabular ML portfolio
  - Classical deep learning: RealMLP (general but relatively slow)
  - Kernel methods: Recursive feature machines (very good for regression, fast on small data)
  - Tabular foundation models: TabICL + extensions (good without tuning, but slow inference)
- Tabular benchmarking with TabArena

Literature I

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Neural Information Processing Systems*, 2022.

David Holzmüller, Leo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. In *Neural Information Processing Systems*, 2024.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *International Conference on Machine Learning*, 2004.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data, February 2025. URL http://arxiv.org/abs/2502.05564. arXiv:2502.05564 [cs].

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383 (6690):1461–1467, 2024. URL https://www.science.org/doi/10.1126/science.adi5639.

TabArena: runtimes