TabPFN for Data-Scarce Industrial Settings

João Machado de Freitas[1], **Alexander Fuchs**[1], **Markus Feuerstein**[2], **Philipp Paller**[2], **Franz Pernkopf**[1]
Christian Doppler Laboratory for Dependable Intelligent Systems in Harsh Environments
SPSC Laboratory - Graz University of Technology[1], RHI Magnesita[2]

Abstract

Tabular foundation models such as *TabPFN v2* perform in-context learning by conditioning on a small labeled support set and a query instance, enabling fast adaptation to heterogeneous tabular *regression* tasks without per-dataset training. Many industrial applications operate in a tiny-sample regime due to cost, and process constraints. We analyze *TabPFN* under extreme label scarcity for regression, positioning it against established tabular baselines and tracing *dataset-size–dependent* predictive performance. Our study analyzes sample sizes from 5 labeled points per task, including an industrial steelmaking regression problem and public benchmarks. In steelmaking, in-process target measurements are rarely feasible, with intermediate targets embedded in delayed end-of-process data. Since data collection is slow and scarce, effective use requires integrating heterogeneous datasets across vessels, processes, and plants.

A central finding is that the *TabPFN* support set size dependency varies widely with dataset quality and information content. While most benchmark tasks achieve satisfactory performance beyond support set sizes of 20, the investigated industrial datasets require at least 100 samples to consistently outperform a naive mean baseline. We discuss implications for deploying in-context tabular models in the low-data regime and show dataset size dependencies for various competitive tabular regression methods.

1 Introduction

Traditional machine learning struggles in real-world industrial settings, where retraining a model for each plant or vessel type is impractical. Here, small data regression tasks are of utmost relevance, but inputs are often a mix of continuous and categorical attributes with task-specific semantics. Especially in industries such as steelmaking, data collection is slow and costly—often taking weeks for a single target value, while in-process measurements are infeasible and intermediate targets remain hidden in end-of-process data. This leads to extremely small datasets due to measurement cost, rarity of events, or process constraints. Moreover, process parameters differ across plants and vessel types, producing many small but related tabular datasets with heterogeneous features. To address this scarcity and variability, all available information must be leveraged across processes.

Fortunately, many industrial tasks can be mapped to tabular data regression problems, a highly researched field, as tabular data underpins decision-making in domains such as healthcare, finance, business operations, and the sciences [8]. In tabular regression, linear models and gradient-boosted decision trees [1, 11]—remain strong baselines because their inductive biases align with common tabular regularities. However, these methods typically require per-dataset training and careful feature handling, which can be cumbersome for task shifts or fragmented data batches.

Foundation models for tabular data aim to provide a single, reusable predictor that can be directly applied across many tasks. The transformer-based Tabular Prior-Fitted Network (TabPFN) family [9, 6, 7] frames prediction as in-context learning. This paradigm is attractive for heterogeneous tabular regression because it reduces per-task training overhead and, in principle, can leverage whatever small support set is available at prediction time [5].

In many industrial regression applications, the central question is not only absolute error on a fixed training split but how prediction quality evolves with the amount of supervision. We therefore study the *sample efficiency* of TabPFN for regression, with a focus on the 5–300 sample regime that frequently arises in practice. Across this range, we compare TabPFN to established tabular baselines and quantify when an in-context model becomes a viable choice. A key observation guiding this paper is, that while TabPFN is relatively sample efficient for the investigated benchmark data, for the investigated industrial datasets, it does *not* reliably exceed an *uninformed* mean regressor at the very smallest sample sizes and only surpasses this baseline once the labeled sample size reaches $\approx$100. This is on average 5 times larger than for the curated benchmark datasets. These findings can guide deployment decisions in low-data regimes, indicating that here performance is mostly decided by data quality rather than model architecture. We provide results for the OpenML CTR23 benchmark [4], including 35 tabular regression datasets and several steelmaking datasets [10, 15].

2 TabPFN

Recent work on tabular foundation models has increasingly focused on *Prior-data Fitted Networks* (PFNs) [9], which approximate Bayesian inference via synthetic training tasks and then performing zero-shot prediction on previously unseen datasets via a single forward pass. By replacing per-dataset optimization and hyperparameter search with in-context inference, PFNs are attractive in settings with scarce data – common in industrial applications. PFNs are trained across tasks under a meta-learning paradigm, require no per-task updates, and provide fast inference that is relatively insensitive to hyperparameters. TabPFN v1 [6] extended PFNs to tabular data by training a transformer on millions of synthetic datasets sampled from generative causal models, encouraging robustness to diverse input-output relations, achieving competitive performance w.r.t. tree-based models and AutoML systems [2]. A persistent obstacle for tabular foundation models is heterogeneity: datasets vary in dimensionality, semantics, and statistical structure, especially in industrial settings where features reflect different plants, processes, vessels, or measurement locations within the same vessel. While TabPFN v1 addressed variable dimensionality by padding attribute vectors to a fixed width, TabPFN v2 [7] introduces a two-dimensional permutation-invariant transformer over samples and features and an attribute tokenizer to handle heterogeneous input spaces and missing values.

During pretraining, however, the number of samples per synthetic task was drawn uniformly up to 2,048 with a fixed validation size of 128, and the total number of table cells per task was capped at 75,000 to control memory usage [7]. As a result, TabPFN v2 is primarily exposed to small- and medium-sized datasets during training. This raises a key question for industrial applications: how well does TabPFN v2 handle highly noisy, heterogeneous, and imputed datasets that may deviate significantly from its training distribution?

3 Data

OpenML CTR23 The OpenML CTR23 benchmark [4, 3, 14] provides a curated collection of 35 tabular regression datasets designed to support reproducible evaluation of machine learning methods. The selection follows explicit inclusion criteria – datasets must contain between 500 and 100,000 observations, fewer than 5,000 one-hot encoded features, no sparse representations, and i.i.d. samples. To validate the benchmark, several standard regression methods are compared in [4, 3, 14].

Vessels The *Vessels* datasets, include two regression tasks derived from real-world steel manufacturing processes. These datasets correspond to two locations from ladle vessels (e.g., slag and metal zones in ladles). Each observation represents a heat described via process parameters, e.g. temperature, chemical additives, and operational settings. The target variable is the refractory wear rate, which is inferred from cumulative end-of-campaign brick-length measurements. To enable supervised learning, we aggregate heats within a campaign using mean and standard deviation, transforming the original multiple-instance regression problem [13] into a campaign-level, tabular regression task with an identical set of numeric and binary process parameters.

4 Experiments

Experimental Setup For both the OpenML CTR23 and Vessels benchmarks, we evaluate model performance in the small-data regime by subsampling training sets of increasing size: from as few as 5 samples up to several hundred (for Vessel)/several thousand (for OpenML CTR23) samples, as well as using the maximum number of available samples per dataset, using 10-fold cross-validation. We evaluate TabPFN v2, and for comparison, we include four widely used regression baselines: *CatBoostRegressor* [12], a gradient-boosted implementation for decision trees that is known for strong performance on tabular data via the CatBoost library. *RandomForestRegressor* and *KNeighborsRegressor* are taken from [11], representing ensemble tree-based and instance-based non-parametric methods, respectively. To provide a lower-bound performance reference, we use an uninformed mean regressor model (*DummyRegressor*) that ignores the input features and generates predictions according to the target mean of the training set. Model performance is reported using mean absolute error (MAE) and the ranks of all representative methods (i.e. lowest MAE of a method corresponds to rank 1), averaged across all folds and random splits.

Results: OpenML CTR23 Across the OpenML CTR23 benchmark, all evaluated methods achieved competitive performance, with results generally consistent with those reported by Fischer et al. [4]. As expected, gradient boosting methods such as CatBoost [12] provided strong baselines, particularly for medium-sized datasets. An evaluation in [16] on classification tasks shows that TabPFN v2 consistently outperforms CatBoost.

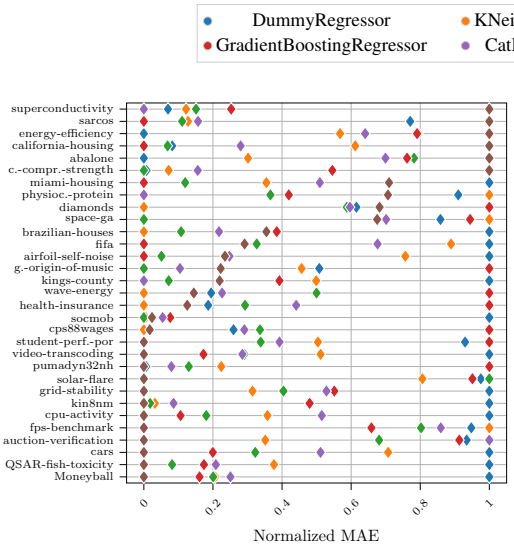

Figure 1: Strip plot showing the normalized MAE for $n = 20$ support samples for all datasets in the OpenML CTR23 benchmark.

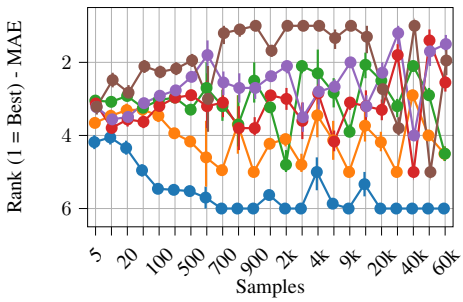

Figure 2: Average rank by support sample size for all datasets in CTR23. Lower rank is better. Error bars represent the 95% confidence interval.

The average ranks of all methods over sample sizes are shown in Figure 2. The confidence interval reflects variability of the estimator's rank on that size. TabPFN v2 is consistently performing best at or near rank 1 for small and medium sample sizes. This is exactly the regime we care about for industrial condition-monitoring datasets and validates TabPFN v2's strength on small, heterogeneous tables. For small data ($\leq$ 600 samples), TabPFN v2 is also best performing, in terms of mean rank, whereas tree ensembles fluctuate more and the KNeighborsRegressor is rarely competitive. Beyond 10,000 samples, the rankings oscillate and there is no definitive winner. The rank of the DummyRegressor is constantly high as expected. TabPFN v2 performs comparably to CatBoost on several datasets, particularly in the small-sample regime up to 100 training points, where its in-context adaptation ability enables fast convergence without retraining. This result supports the hypothesis that meta-learned priors are beneficial for tabular regression tasks with scarce data. The MAE for all OpenML datasets normalized to 0 (1) for the best (worst) model for $n = 20$ support

samples is shown in Fig. 1. TabPFN achieves competitive normalized MAE with fewer samples (see also ranking in Fig 2 for $n = 20$), suggesting that sample efficiency is a key advantage of this approach.

Results: Vessels The MAE for different sample sizes of the ladle slag and metal dataset is shown in Figure 3 and 4, respectively. For small sample regimes ($\leq$ 100 samples) all methods achieved similar MAE, with performance close to the uninformed mean regressor. This result is consistent across both ladle vessels – the metal and the slag zone – suggesting that model selection has limited impact given the inherent noise, missingness, and heterogeneity of the data. For larger sample regimes ($\geq$ 200 samples) all *informed* methods significantly outperform the uninformed mean regressor. These methods obtain similar MAE performance with a slight preference for TabPFN for the ladle slag dataset. Furthermore, the curves over the sample sizes of each method show that performance does not consistently improve with more samples – in some cases, additional data even worsens generalization due to accumulated noise. These findings have the value that they highlight that standard regression models saturate quickly on this benchmark, implying that improvements are more likely to come from better data preprocessing, feature engineering, or domain-informed modeling rather than from selecting a different tabular learner.

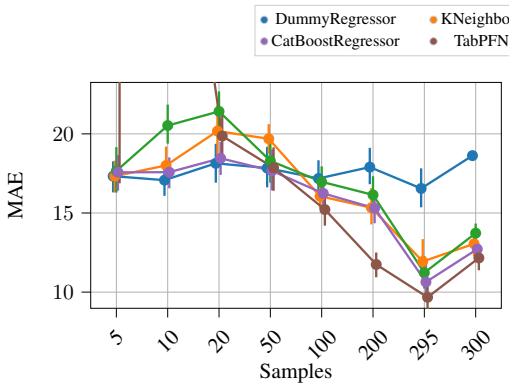

Figure 3: MAE for different sample sizes of ladle slag dataset.

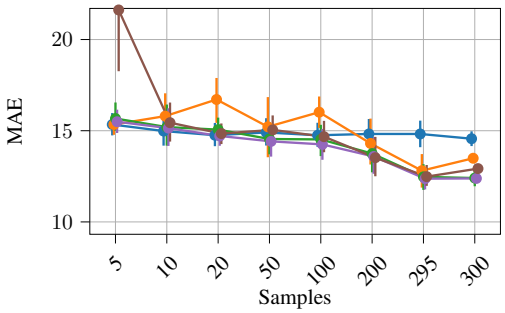

Figure 4: MAE for different sample sizes of ladle metal dataset.

These results on industrial data suggest that model performance is fundamentally constrained by data quality rather than model capacity. They underline the importance of (i) improved measurement strategies during campaigns, (ii) better alignment of process parameters across plants, and (iii) integration of physical priors to regularize learning in this extreme data-scarcity setting.

5 Conclusion

We evaluated the applicability and support set dependency of TabPFNv2, to industrial prediction problems, using the CTR23 benchmark and condition monitoring datasets from steel manufacturing as representative case. On the curated OpenML CTR23 benchmark, TabPFN is relatively sample efficient and often competitive against the selected baseline methods. Here, it requires on average $\geq$ 50 support samples to consistently outperform alternatives. On a dataset-level we also see variability in performance, indicating uneven information density and data quality across tasks within the benchmark. On the industrial Vessels benchmark, TabPFN only begins to reliably beat an uninformed regressor beyond $\approx$100 samples, pointing to lower per-sample information driven by label sparsity, measurement noise, and heterogeneity—rather than model capacity.

These results suggest that whether TabPFNs are applicable in industry depends less on the model architecture and more on the quality and information content of the support set. Our break-even points at $\approx$20 samples on CTR23 and $\approx$100 on Vessels offer guidance for deploying in-context TabPFNs to industrial data and set realistic expectations: while TabPFNs enable rapid, retraining-free inference, their reliability depends on increasing the effective information content of the context rather than replacing models.

Acknowledgments and Disclosure of Funding

The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged. Furthermore, the research was funded by RHI Magnesita.

References

[1] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2011. In press.

[2] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data, 2020.

[3] Matthias Feurer, Jan N. van Rijn, Arlind Kadra, Pieter Gijsbers, Neeratyoy Mallik, Sahithya Ravi, Andreas Mueller, Joaquin Vanschoren, and Frank Hutter. OpenML-Python: an extensible Python API for OpenML. *arXiv*, 1911.02490, 2019.

[4] Sebastian Fischer, Liana Harutyunyan, Matthias Feurer, and Bernd Bischl. OpenML-CTR23 – a curated tabular regression benchmarking suite. In *AutoML Workshop Track*, 2023. Available as preprint: OpenReview ID HebAOoMm94.

[5] Josh Gardner, Juan C Perdomo, and Ludwig Schmidt. Large scale transfer learning for tabular data via language modeling. *Advances in Neural Information Processing Systems*, 37:45155–45205, 2024.

[6] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

[7] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[8] Stephanie L Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, 2020.

[9] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

[10] Nikolaus Mutsam, Alexander Fuchs, Fabio Ziegler, and Franz Pernkopf. Data-scarce condition modeling requires model-based prior regularization. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6695–6699, 2024.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *NeurIPS*, pages 6639–6649, 2018.

[13] Soumya Ray and David Page. Multiple instance regression. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 425–432, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[14] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

[15] Andreas Viertauer, Nikolaus Mutsam, Franz Pernkopf, Andreas Gantner, Georg Grimm, Waltraud Winkler, Gregor Lammer, and Alexander Ratz. Refractory condition monitoring an dlifetime prognosis for rh degasser. In *AISTech – The Iron & Steel Technology Conference and Exposition*, 2019.

[16] Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at tabpfn v2: Understanding its strengths and extending its capabilities, 2025.