

# Closing the Context Gap: Activation Alignment for Tabular In-Context Learning

Yoel Zeldes  
Independent Researcher

yoel.zeldes@gmail.com

## Abstract

Tabular foundation models perform in-context learning (ICL) by conditioning predictions on labeled training examples provided as context. Unlike traditional models that separate training from inference, these models must process all training examples in every forward pass, making each prediction expensive. Restricting the number of training examples reduces this cost but substantially degrades performance. Instead of discarding context, we propose *activation alignment*, a method that leverages the full context to teach a model how to behave when seeing only a subset. This is achieved by training a lightweight linear transformation on synthetic unlabeled data to map the intermediate activations of a data-constrained “student” (using partial context) toward those of a full-context “teacher” (using all data). Training the aligner requires no GPU and converges in seconds to minutes on commodity hardware. We evaluate on 38 classification datasets from the TabArena benchmark using the leading two tabular foundation models, TabPFN-3 and TabFM. Across all context budgets, the aligned student yields broad, statistically significant improvements over the unaligned baseline for both models. In low-data regimes, alignment recovers nearly half of the teacher’s predictive advantage. The method provides a practical, low-overhead approach to achieving the inference speed of compact contexts while closing a significant fraction of the performance gap to the full-context teacher.

## 1 Introduction

Tabular foundation models have become the state of the art for structured data prediction (Hollmann et al., 2025). Models such as TabPFN-3 (Grinsztajn et al., 2026) and TabFM (Google Research, 2026) perform in-context learning (ICL): given a set of labeled training examples as context and an unlabeled query, they produce a prediction in a single forward pass without weight updates. This paradigm enables rapid adaptation to new datasets and eliminates the need for task-specific training.

However, the predictive quality of ICL depends heavily on the number of context examples. When the context is small (due to limited data availability, hardware memory constraints, or computational cost of the quadratic attention mechanism), performance degrades substantially. Garg et al. (2022) showed that transformers implementing ICL improve with the number of in-context examples, and tabular foundation models exhibit a similar dependence on context size in practice (Qu et al., 2025; Ma et al., 2025), making sample efficiency a central bottleneck for practical deployment.

The root cause is architectural: unlike tree-based models, where training and inference are decoupled and prediction is immediate, tabular foundation models must carry all training examples through the forward pass at every prediction. Although intermediate key-value states could be cached, doing so requires large device memory. Existing approaches reduce this inference cost by compressing the context into compact latent tokens (TACO; Zabërgja et al., 2026) or by pruning redundant transformer layers (TACTICL; Koshil et al., 2026). Both methods are effective but demand substantial GPU training.

Inspired by prior work on activation steering in large language models (Turner et al., 2025; Li et al., 2023; Todd et al., 2024) and feature-based knowledge distillation (Romero et al., 2015), we take a complementary

---

approach. Rather than compressing the input or the architecture, we ask: can a model that has seen the full context teach a data-constrained copy of itself how to behave, by correcting its internal representations?

In this work, we propose **activation alignment**, a framework that fits a lightweight linear module to project the intermediate activations of a data-constrained student model toward those of a full-context teacher. Similar to TACO, the aligned student operates with a reduced number of input tokens compared to the full dataset. However, the aligner is drastically cheaper to train: a single linear layer fitted on pre-extracted activations, requiring no GPU and converging within seconds to minutes on commodity hardware. At inference time, the aligner corrects the student’s hidden states at a single transformer layer without requiring any model fine-tuning.

We systematically evaluate activation alignment across 38 classification tasks from the TabArena benchmark (Erickson et al., 2026), a standardized evaluation suite of diverse tabular tasks, using TabPFN-3 (Grinsztajn et al., 2026) and TabFM (Google Research, 2026). Across both models and all evaluated context budgets, the aligned student outperforms the unaligned baseline in 81% of cases, with statistically significant gains. With only 10% of the training context, the aligned models achieve predictive parity with unaligned students provided two to four times as much data. Because activation alignment modifies internal representations rather than inputs or architecture, it is orthogonal to existing compression and context-selection methods; we discuss how the two approaches could be composed in Section 6.<sup>1</sup>

## 2 Related Work

### 2.1 Tabular Foundation Models and In-Context Learning

Tabular foundation models perform prediction via in-context learning, processing labeled training examples and test queries in a single forward pass (Hollmann et al., 2022; Grinsztajn et al., 2026). TabPFN is a Prior-Data Fitted Network (Müller et al., 2022) trained on synthetic data from structural causal models. While TabPFN v2 employed feature×item tokenization with alternating attention (Hollmann et al., 2025), TabPFN-3 returned to processing tabular data with a single unified token per example, operating across 24 ICL blocks with multi-estimator ensembling. TabFM similarly adopts row-level tokenization with 24 ICL transformer blocks. While tree-based methods such as gradient-boosted decision trees have long outperformed standard deep learning on tabular data (Grinsztajn et al., 2022), tabular foundation models have emerged that frequently rival or surpass them (McElfresh et al., 2023). Nonetheless, model performance depends heavily on the number of context examples (Hollmann et al., 2025; Qu et al., 2025; Ma et al., 2025).

### 2.2 Compressing Tabular Foundation Models

Because the computational cost and memory footprint of tabular foundation model inference scale with context size, several lines of research explore model and context compression. On the context dimension, TACO (Zabërgja et al., 2026) compresses in-context training examples into compact latent vectors. In a related vein, TuneTables (Feuer et al., 2024) optimizes the context itself: it learns, via gradient descent through the frozen backbone, a small set of soft prompt embeddings that distill a large labeled training set into a compact learned context. In contrast to our aligner, these methods modify the model’s input rather than its internal states and require GPU backpropagation through the model. Along the architectural dimension, TACTICL (Koshil et al., 2026) exploits depth-wise redundancy documented by Gromov et al. (2024) by pruning transformer layers and substituting them with lightweight adapters, while MotherNet (Mueller et al., 2025) trains hypernetworks to synthesize small feed-forward networks directly. A complementary strategy acts on context *selection* rather than compression: LoCalPFN (Thomas et al., 2024) retrieves the  $k$  nearest neighbors of each query from the training set to serve as a per-query local context, and fine-tunes TabPFN end-to-end with those retrieved neighbors in context.

Our approach shares with TACO the objective of operating over reduced token counts at inference time while retaining the full data’s predictive quality. However, existing methods demand computationally intensive offline procedures, such as joint end-to-end pretraining of both compressor and predictor networks (TACO),

---

<sup>1</sup>Code to reproduce all experiments is available at <https://github.com/yoel-zeldes/tabalign>.

training layer-replacement adapters with end-to-end fine-tuning (TACTICL), or expensive hypernetwork optimization (MotherNet). In contrast, our linear aligner is trained directly on pre-extracted representations without backpropagating through the model, converging within seconds to minutes on ordinary CPU cores.

## 2.3 Feature-Based Knowledge Distillation

To transfer knowledge from compute- or data-rich regimes without retraining the underlying model, knowledge distillation offers a well-established framework. Hinton et al. (2015) introduced matching output-level soft probabilities. Feature-based methods extend this to intermediate representations: FitNets (Romero et al., 2015) trains the student to match teacher hidden activations via a learned regressor; attention transfer (Zagoruyko & Komodakis, 2017) aligns attention maps; and contrastive representation distillation (Tian et al., 2022) uses contrastive objectives for structural knowledge transfer.

These methods distill knowledge across architectures, from a large teacher to a smaller student. In contrast, our work operates within the same frozen architecture, transferring information across data regimes (full context to reduced context) using synthetic unlabeled queries. A further distinction is procedural: in feature-based distillation the learned regressor is training-time scaffolding, discarded once the student’s weights are updated, whereas our aligner leaves all weights frozen and is itself the deployed inference-time intervention.

## 2.4 Activation Steering

While distillation provides the objective, the feasibility of modifying intermediate states at test time without retraining is supported by recent advances in representation engineering and activation steering. A growing body of literature demonstrates that hidden states across transformer layers can be directly manipulated during inference to guide downstream outputs. Representation Engineering (Zou et al., 2023) and related activation steering methods demonstrate that high-level concepts are encoded in population-level activations, enabling techniques like Activation Addition (Turner et al., 2025) and Inference-Time Intervention (Li et al., 2023) to shift hidden states along desired directions, such as task intent or truthfulness. Similarly, Arditì et al. (2024) showed that refusal mechanisms in language models are mediated by a single linear direction in the residual stream, while function vectors (Todd et al., 2024) revealed that entire in-context learning tasks can be compactly localized within intermediate attention head outputs.

Tuned Lens (Belrose et al., 2023) is particularly related to our approach: it trains an affine probe at each layer of a frozen pretrained model to decode intermediate hidden states into the model’s final output distribution. Like our aligner, Tuned Lens learns a linear transformation of a frozen model’s intermediate activations. The key difference is in purpose: Tuned Lens is an interpretability tool that reads predictions from intermediate representations, while our aligner writes corrected representations back into the forward pass to improve downstream predictions.

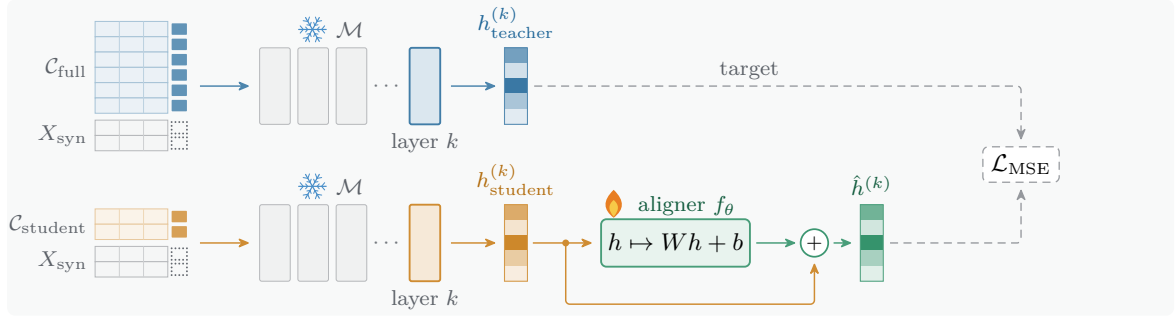
# 3 Method

## 3.1 Problem Formulation

Let  $\mathcal{M}$  be a tabular foundation model that performs in-context learning: given a labeled context set  $\mathcal{C} = \{(x_i, y_i)\}_{i=1}^N$  and a query  $x_q$ , the model produces a prediction  $\hat{y}_q = \mathcal{M}(x_q \mid \mathcal{C})$  in a single forward pass without parameter updates. We consider the setting where a *teacher* has access to the full context  $\mathcal{C}_{\text{full}}$  of size  $N_{\text{full}}$ , while a *student* is restricted to a subset  $\mathcal{C}_{\text{student}} \subset \mathcal{C}_{\text{full}}$  of size  $N_{\text{student}} = \lfloor \alpha \cdot N_{\text{full}} \rfloor$ , where  $\alpha \in (0, 1)$  is the student fraction. The student context  $\mathcal{C}_{\text{student}}$  is a stratified random subset of  $\mathcal{C}_{\text{full}}$ , preserving the label distribution.

Both teacher and student use the same frozen model  $\mathcal{M}$ . The teacher and student differ only in their context: the student’s intermediate representations carry less task information because they are conditioned on fewer labeled examples. We aim to learn a lightweight transformation that maps the student’s intermediate activations toward the teacher’s, thereby recovering some of the predictive quality lost from the reduced context.

**(a) Offline: fit the aligner**



**(b) Inference: correct the activation**

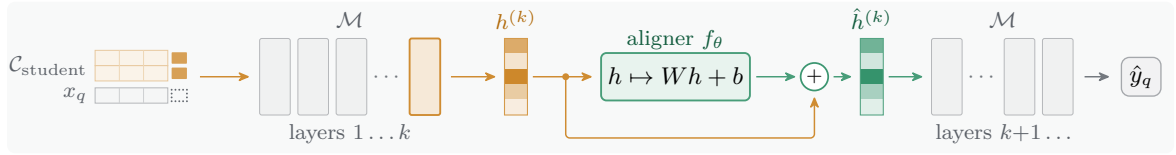


Figure 1: **(a)** Unlabeled synthetic queries  $X_{\text{syn}}$  are appended to the full context (teacher) and to the reduced context (student), and both are passed through the same frozen model  $\mathcal{M}$ . A linear aligner  $f_\theta$  is fit by MSE so that the corrected student activation matches the teacher’s at layer  $k$ . **(b)** At test time, the trained aligner is applied to the student’s layer- $k$  activations.

### 3.2 Pipeline Overview

The method consists of four stages (Figure 1):

1. **Synthetic data generation:** Generate unlabeled query vectors  $X_{\text{syn}}$  from the feature distribution, requiring no additional labels.
2. **Activation extraction:** Pass  $X_{\text{syn}}$  through  $\mathcal{M}$  under both teacher and student contexts, extracting intermediate activations at a chosen layer  $k$ .
3. **Aligner training:** Fit a linear model to map student activations to teacher activations using MSE loss on the extracted activation pairs.
4. **Inference-time alignment:** Apply the trained aligner to transform the student’s activations at layer  $k$  during the forward pass.

### 3.3 Synthetic Data Generation

To train the aligner without requiring additional labeled data, we generate  $N_{\text{syn}}$  synthetic query vectors  $X_{\text{syn}} \in \mathbb{R}^{N_{\text{syn}} \times m}$  (where  $m$  is the number of features) by using TabPFN’s unsupervised generative capabilities to model the training features  $X_{\text{train}}$  and sampling from the learned joint multivariate distribution. This captures feature correlations present in the real data. In our experiments, we use  $N_{\text{syn}} = 1,000$  synthetic samples. For TabFM, we use the same TabPFN-generated synthetic queries; as the aligner operates on model-internal activations, the synthetic input distribution serves only to probe the representation space and is not model-specific.

### 3.4 Aligner Training

The aligner  $f_\theta$  is a linear model  $f_\theta(h) = Wh + b$ , where  $W \in \mathbb{R}^{d \times d}$  and  $b \in \mathbb{R}^d$ , operating on intermediate representations of hidden dimension  $d$  at a designated layer  $k$ .

**Residual formulation.** Rather than predicting the teacher activation directly, the aligner predicts the residual  $\Delta h_i = h_{\text{teacher},i}^{(k)} - h_{\text{student},i}^{(k)}$ , where  $h_{\text{teacher},i}^{(k)}, h_{\text{student},i}^{(k)} \in \mathbb{R}^d$  denote the layer- $k$  activations of the  $i$ -th synthetic query under the full context  $\mathcal{C}_{\text{full}}$  and the student context  $\mathcal{C}_{\text{student}}$ , respectively, yielding  $N_{\text{syn}}$  activation pairs. The weights are initialized with near-zero values, so that the aligner initially outputs near-zero corrections, preserving the student’s original representations at the start of training. At inference, the aligned activation is  $\hat{h}^{(k)} = h_{\text{student}}^{(k)} + f_{\theta}(h_{\text{student}}^{(k)})$ .

**Layer selection.** For both TabPFN and TabFM (each comprising 24 layers), we extract and align activations at the final layer.

**Training objective.** The aligner is trained to minimize MSE loss:

$$\mathcal{L} = \frac{1}{N_{\text{syn}}} \sum_{i=1}^{N_{\text{syn}}} \left\| f_{\theta}(h_{\text{student},i}^{(k)}) - \Delta h_i \right\|_2^2 \quad (1)$$

**Optimization.** We use AdamW with a learning rate of  $10^{-3}$  for TabPFN and  $10^{-4}$  for TabFM, and early stopping with a patience of 10 epochs on a held-out validation split (80/20 random split of the activation pairs).

**Multi-estimator models.** We use an 8-estimator ensemble for TabPFN and a single estimator for TabFM. For TabPFN, a separate aligner is trained independently for each estimator.

**Shared preprocessing.** Tabular foundation models rely on data-dependent preprocessing pipelines, such as quantile transforms, categorical encodings, and randomized feature subsets or permutations per ensemble member. To ensure consistent representations, the student inherits the teacher’s fitted preprocessors while conditioning its in-context learning strictly on the reduced subset  $\mathcal{C}_{\text{student}}$ . This guarantees that each student estimator operates in the same input representation space as its teacher counterpart, avoiding representation drift caused by differing random permutations, mismatched category codes, or unstable quantile estimates on small sample subsets, which would introduce arbitrary distortions and render the alignment task ineffective.

## 4 Experimental Setup

**Benchmark.** We evaluate on 38 binary and multiclass classification datasets from the TabArena-v0.1 benchmark suite (Erickson et al., 2026).

**Student fractions.** For each dataset, we vary the student fraction  $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ , where  $N_{\text{student}} = \lfloor \alpha \cdot N_{\text{full}} \rfloor$ . We focus our primary analysis on context reductions of  $2\times$  to  $10\times$  ( $\alpha \leq 0.5$ ), where context truncation delivers the most critical computational and memory savings. For completeness, evaluations across the full spectrum  $\alpha \in \{0.1, \dots, 0.9\}$  are provided in Table 1, with consistent conclusions.

**Repeats.** Each experiment is repeated 5 times using the predefined train/test splits attached to each dataset’s OpenML task (Vanschoren et al., 2014), to account for variability.

**Metric.** We use ROC-AUC for binary classification datasets and negative log-loss for multiclass datasets (higher is better).

**Baselines.** We compare four conditions at each student fraction:

1. **Baseline student:**  $\mathcal{M}$  with context  $\mathcal{C}_{\text{student}}$  (no alignment).
2. **Aligned student:**  $\mathcal{M}$  with context  $\mathcal{C}_{\text{student}}$  and the trained aligner applied at the last layer.
3. **Teacher:**  $\mathcal{M}$  with the full context  $\mathcal{C}_{\text{full}}$  (upper bound).

- 
4. **XGBoost:** XGBoost (Chen & Guestrin, 2016) trained on  $\mathcal{C}_{\text{full}}$  using the standard default hyperparameters established in Gorishniy et al. (2021). This represents the standard practical alternative when full-context transformer inference is constrained by compute or memory, as tree-based models can utilize the entire dataset without incurring higher inference latency.

**Effective sample fraction.** To quantify the alignment benefit, we compute an effective sample fraction  $E_\alpha$ : the student fraction at which the unaligned student would need to operate to achieve the same metric score as the aligned student at fraction  $\alpha$ . This is obtained by linear interpolation between baseline scores at neighboring fractions. An aligned student with  $E_\alpha > \alpha$  is effectively operating as if it had more data.

**Models tested.** We evaluate on two tabular foundation models: TabPFN-3 with an 8-estimator ensemble and TabFM with a single estimator.

**Teacher superiority condition.** We evaluate alignment metrics on instances where the teacher outperforms the unaligned baseline. When a student already matches or surpasses the full-context teacher (e.g., due to split variance on small datasets), distilling teacher representations is unmotivated. Across all  $38 \times 5 = 190$  evaluation slices (38 datasets across 5 context fractions), this condition excludes only 5 slices for TabPFN and 2 for TabFM.

## 5 Results

We evaluate activation alignment across three central empirical questions: (1) Does aligning intermediate representations systematically improve downstream performance over the unaligned baseline? (2) How does the effective data gain behave across varying context constraints? (3) Can a data-constrained foundation model with alignment compete with or surpass a strong tree-based model trained on the entire dataset?

### 5.1 Alignment Substantially Improves Over Baseline Student

Figure 2 presents pairwise win rates with Wilcoxon signed-rank tests against both the unaligned baseline student and full-data XGBoost, and Figure 3 illustrates the scaling curves and distillation dynamics (detailed per-dataset benchmarks across all context budgets  $\alpha \in \{0.1, \dots, 0.9\}$  are reported in Table 1).

Across diverse tabular tasks, intermediate activation alignment yields broad, statistically significant performance gains over the unaligned baseline student (Figure 2). Statistical significance is evaluated via a two-sided paired Wilcoxon signed-rank test across datasets (Demšar, 2006): for each individual context fraction  $\alpha$ , the test pairs the 5-repeat mean scores of aligned models against the respective comparator (the unaligned baseline student or full-context XGBoost) across all evaluated datasets. To evaluate overall statistical significance across all context budgets without pseudoreplication (i.e., violating the independence assumption by treating multiple context fractions of the same dataset as independent observations), the test pairs the per-dataset mean scores averaged across all context fractions. Against the unaligned baseline student, TabPFN achieves an 82.2% win rate over all evaluated (dataset, context fraction) pairs, with performance improvements being statistically significant overall ( $p < 0.0001$ ) and across every individual context fraction. For TabFM, the aligned student achieves a 79.8% overall win rate against the baseline student ( $p < 0.0001$ ), maintaining statistically significant gains across every evaluated context fraction.

### 5.2 Sample Efficiency and Distillation Gains

To understand how the aligner modulates task representations under varying data constraints, we examine the effective sample fraction  $E_\alpha$  and the proportion of the teacher–student gap recovered (Figure 3).

**Data amplification in the low-data regime.** In extreme low-data settings ( $\alpha = 0.1$ , corresponding to 10% of available training points), the aligned student achieves a median effective sample fraction of  $E_{0.1} = 0.235$  for TabPFN and 0.200 for TabFM (with mean  $E_{0.1} \approx 0.33$  across both architectures, as can be seen in Table 1). Without processing additional in-context tokens during inference, the aligned student

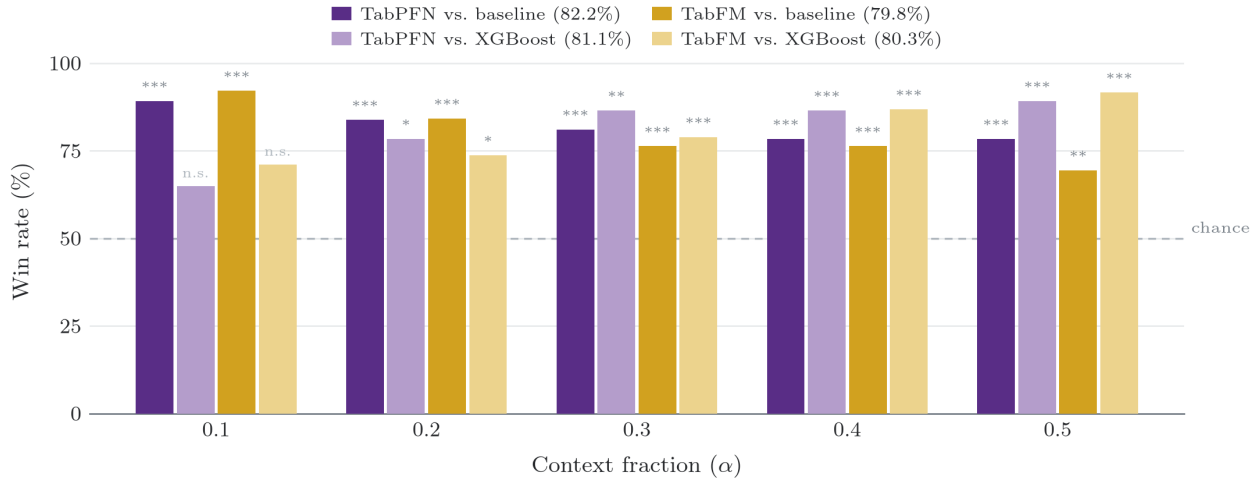


Figure 2: Pairwise win rates of the aligned student over the unaligned baseline student and full-training-budget XGBoost across context fractions  $\alpha \in \{0.1, \dots, 0.5\}$  on 38 TabArena classification datasets (5 repeats) for TabPFN and TabFM. Overall win rates across all context budgets are reported in the legend (82.2% vs. base and 81.1% vs. XGBoost for TabPFN; 79.8% vs. base and 80.3% vs. XGBoost for TabFM). Statistical significance annotations above each bar indicate two-sided paired Wilcoxon signed-rank test results against the respective comparator ( $***p < 0.001$ ,  $**p < 0.01$ ,  $*p < 0.05$ ; n.s.: not significant).

attains predictive accuracy comparable to an unaligned model provided with two to four times as many labeled examples. Crucially, median sample efficiency remains strictly above parity across all evaluated context budgets.

**Gap closure across context budgets.** Examining the relative gap closed reveals consistent recovery of the teacher’s predictive advantage: across all evaluated context fractions, the median gap closed remains above 34% for TabPFN and above 24% for TabFM. While alignment provides non-trivial benefits across the entire context spectrum, it is particularly effective in lower-data regimes: at  $\alpha = 0.1$ , alignment recovers a median of 48.3% and 44.7% of the teacher–student performance deficit for TabPFN and TabFM, respectively.

### 5.3 Comparison with Full-Data XGBoost

A central practical question for tabular foundation models is whether context-constrained in-context learning can match standard tree-based models trained on the complete dataset. While XGBoost had access to 100% of the training data ( $\mathcal{C}_{\text{full}}$ ), the aligned student was restricted to subsets between 10% and 50% of the context examples.

Despite this substantial data handicap, the aligned student broadly outperforms full-training-budget XGBoost across both architectures (Figure 2). Overall, the aligned student beats full-data XGBoost in 81.1% of conditions for TabPFN and 80.3% for TabFM ( $p < 0.0001$  overall for both models), maintaining statistically significant advantages across all context budgets  $\alpha \geq 0.2$ . In a 3-way ranking (aligned student vs. baseline student vs. full-data XGBoost), the aligned student maintains the top average rank across all context budgets: 1.37 for TabPFN and 1.40 for TabFM, compared to 2.19 and 2.13 for the baseline student and 2.44 and 2.47 for XGBoost (Figure 4). These results demonstrate that combining tabular ICL models with activation alignment can surpass fully trained gradient boosting models even under strict context budgets.

## 6 Limitations and Future Work

### 6.1 Limitations

Several limitations bound the scope of these findings:

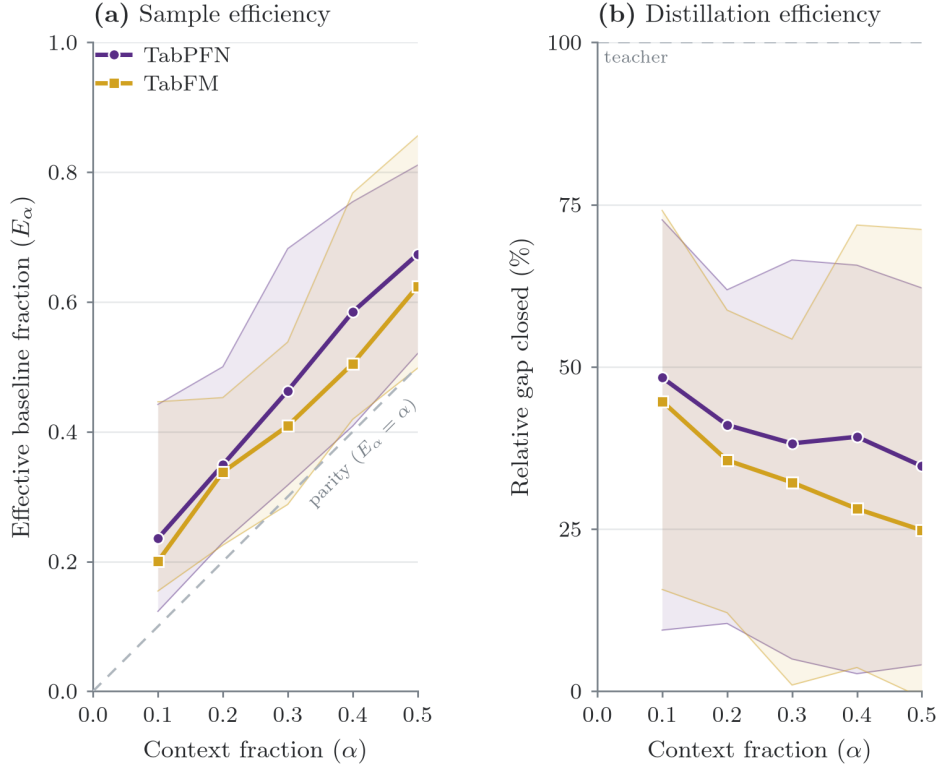


Figure 3: Median scaling and distillation efficiency across 38 TabArena classification datasets for TabPFN and TabFM, averaged over 5 repeats. Shaded bands denote the 25th–75th percentiles across datasets. **(a) Sample Efficiency:** Median effective baseline context fraction  $E_\alpha$  achieved by the aligned student as a function of the actual context budget  $\alpha$ . Both models achieve positive data gains across all context fractions. **(b) Distillation Efficiency:** Median relative gap closed between baseline student and full-context teacher across context fractions  $\alpha$ , showing that alignment closes a significant fraction of the gap across all context budgets.

1. **Classification only.** While our evaluation focused on classification tasks, the aligner operates entirely at the level of intermediate transformer activations and uses unlabeled synthetic queries, with no classification-specific assumptions. We therefore expect the method to translate directly to regression settings, though this has yet to be empirically verified.
2. **Model coverage.** Our empirical evaluation was conducted on two representative tabular foundation models (TabPFN-3 and TabFM). We expect that activation alignment generalizes to other tabular in-context learning architectures, but validating this across a broader family of foundation models remains an important empirical step.
3. **Not universally beneficial.** Five of the 38 datasets (MIC, SDSS17, credit\_card\_clients\_default, kddcup09\_appetency, and polish\_companies\_bankruptcy) show a negative mean effective-sample gain for both models; Table 1 provides the full per-dataset breakdown. We do not currently have a characterization that separates these datasets from those that benefit. Identifying such properties remains an open question and would allow practitioners to determine in advance whether alignment is worth applying to a given task.

## 6.2 Future Directions

The results suggest several directions for future work:

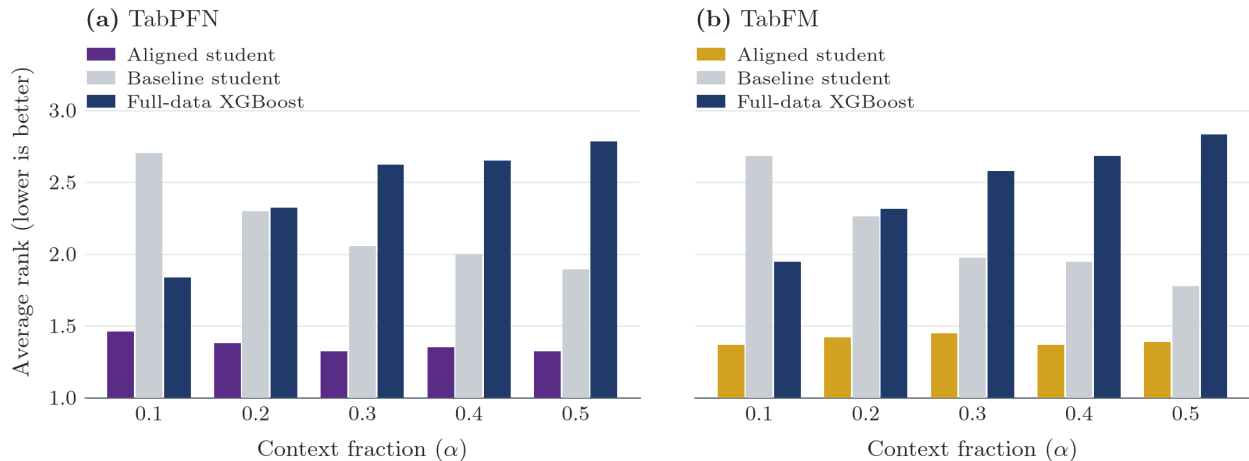


Figure 4: 3-way average rank (lower is better) across 38 TabArena classification datasets for (a) TabPFN and (b) TabFM, averaged over 5 repeats. At each context fraction  $\alpha \in \{0.1, \dots, 0.5\}$ , bars compare the aligned student, unaligned baseline student, and full-context XGBoost. Across all context budgets, the aligned student maintains the top average ranking.

- **Expressive aligners and training objectives.** Future work could explore non-linear aligners (e.g., multi-layer perceptrons), adaptive layer selection, intervening across multiple layers simultaneously, systematic hyperparameter optimization, and alternative training objectives such as incorporating an end-to-end distillation loss over output logits.
- **Combination with compression and context selection.** Activation alignment could be applied to models compressed via context compression (e.g., TACO) or layer pruning (e.g., TACTICL). Because compression inevitably perturbs intermediate activations away from those of the uncompressed model, learning an aligner to map the compressed model’s representations back toward the full-capacity teacher could recover performance lost to compression while retaining inference speedups. The same composition applies to retrieval-based context selection (e.g., Thomas et al., 2024): our student’s context is a stratified random subset and we do not compare against selection strategies, but an aligner could equally correct the representations of a student whose context was chosen by retrieval.
- **Mechanistic analysis.** While our results establish that a linear correction is sufficient to recover a substantial share of the teacher–student gap, they do not characterize the geometric structure of the representation displacement. Future mechanistic work analyzing intermediate representation spaces across architectures could clarify why linear alignment is effective, whether non-linear structure remains uncaptured, and why different models exhibit differing degrees of alignment gains.
- **Aligners for genuinely small datasets.** Currently, the aligner requires offline access to the full dataset to extract teacher activations, which addresses inference-time compute and memory bottlenecks but does not apply when only a small dataset exists to begin with. The framework could be extended to genuinely low-data regimes by conditioning the aligner on summary statistics or domain priors (e.g., population moments that differ from small-sample estimates). An aligner could be trained across diverse external datasets where abundant data is available, learning to map student activations and summary statistics to teacher representations. At test time, this model could then align representations on genuinely small, unseen datasets where additional examples do not exist.

---

## 7 Conclusion

In this work, we introduce activation alignment to address the sample efficiency bottleneck in tabular in-context learning. By projecting intermediate hidden states of data-limited student models toward those elicited under full context at a single layer, our approach enables substantially higher predictive performance from small context budgets. The aligner is parameterized as an affine transformation optimized via MSE over synthetic data, requiring no external labels, running on commodity hardware, and leaving the underlying backbone frozen.

Across 38 classification benchmarks from TabArena evaluated on both TabPFN-3 and TabFM, activation alignment yields broad, statistically significant gains over unaligned baselines. Gains are particularly striking in low-data regimes: with only 10% of training examples as context, aligned students match the predictive accuracy of unaligned models provided two to four times as much data. Furthermore, aligned students operating on partial contexts systematically outperform full-data gradient-boosted trees in over 80% of evaluations.

More broadly, our findings demonstrate that intermediate representations in tabular foundation models are readily steerable toward high-data regimes using simple linear mappings. Activation alignment offers a practical, model-agnostic technique for overcoming memory and context limitations in real-world tabular machine learning.

## Acknowledgments

The author thanks Alan Arazi (Technion – Israel Institute of Technology) for valuable discussions and feedback.

## References

- Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083. Curran Associates, Inc., 2024. doi: 10.52202/079017-4322. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf).
- Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112, 2023. URL <https://doi.org/10.48550/arXiv.2303.08112>.
- Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL <https://doi.org/10.1145/2939672.2939785>.
- Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL <http://jmlr.org/papers/v7/demsar06a.html>.
- Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for machine learning on tabular data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL <https://openreview.net/forum?id=jZqCqpCLdU>.
- Benjamin Feuer, Robin Tibor Schirrmester, Valeriia Cherepanova, Chinmay Hegde, Frank Hutter, Micah Goldblum, Niv Cohen, and Colin White. TuneTables: Context optimization for scalable prior-data fitted networks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 83430–83464. Curran Associates, Inc.,

- 
2024. doi: 10.52202/079017-2654. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/97dc07f1253ab33ee514f395a82fa7cc-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/97dc07f1253ab33ee514f395a82fa7cc-Paper-Conference.pdf).
- Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=f1NZJ2e0et>.
- Google Research. TabFM: A zero-shot foundation model for tabular data. 2026. URL <https://research.google/blog/introducing-tabfm-a-zero-shot-foundation-model-for-tabular-data/>.
- Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL [https://openreview.net/forum?id=i\\_Q1yrOegLY](https://openreview.net/forum?id=i_Q1yrOegLY).
- Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL [https://openreview.net/forum?id=Fp7\\_\\_phQszn](https://openreview.net/forum?id=Fp7__phQszn).
- Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Elliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-3: Technical report, 2026. URL <https://arxiv.org/abs/2605.13986>.
- Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Dan Roberts. The unreasonable ineffectiveness of the deeper layers. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL <https://openreview.net/forum?id=jwhPERqvds>.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL <https://arxiv.org/abs/1503.02531>.
- Noah Hollmann, Samuel Müller, Katharina Eggersperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *NeurIPS 2022 First Table Representation Workshop*, 2022. URL <https://openreview.net/forum?id=eu9fVjVasr4>.
- Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeyer, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 01 2025. doi: 10.1038/s41586-024-08328-6.
- Mykhailo Koshil, Matthias Feurer, and Katharina Eggersperger. TACTICL: Task-aware compression of tabular ICL models. In *AutoML 2026*, number arXiv:2608.10837. arXiv, August 2026. URL <https://arxiv.org/abs/2608.10837>.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=pIZxE0ZCId>.

- 
- Duncan C. McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=CjVdXey4zT>.
- Andreas C Mueller, Carlo A Curino, and Raghu Ramakrishnan. MotherNet: Fast training and inference via hyper-network transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=6H4jRWKFc3>.
- Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=KSugKcbNf9>.
- Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: a tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML’25. JMLR.org, 2025.
- Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets, 2015. URL <https://arxiv.org/abs/1412.6550>.
- Valentin Thomas, Junwei Ma, Rasa Hosseinzadeh, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony Caterini. Retrieval & fine-tuning for in-context tabular models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 108439–108467. Curran Associates, Inc., 2024. doi: 10.52202/079017-3442. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/c40daf14d7a6469e65116507c21faeb7-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/c40daf14d7a6469e65116507c21faeb7-Paper-Conference.pdf).
- Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation, 2022. URL <https://arxiv.org/abs/1910.10699>.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=AwyxtyMwaG>. arXiv:2310.15213.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2025. URL <https://openreview.net/forum?id=2XBPdPIcFK>.
- Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. *SIGKDD Explor. Newsl.*, 15(2):49–60, June 2014. ISSN 1931-0145. doi: 10.1145/2641190.2641198. URL <https://doi.org/10.1145/2641190.2641198>.
- Guri Zabërgja, Rafiq Kamel, Arlind Kadra, Christian Frey, and Josif Grabocka. End-to-end compression for tabular foundation models. In *Forty-third International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=84mfkGDxYh>.
- Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. URL [https://openreview.net/forum?id=Sks9\\_ajex](https://openreview.net/forum?id=Sks9_ajex).
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023.

---

## A Per-Dataset Sample Efficiency Breakdown

Table 1 reports the detailed per-dataset effective baseline sample fraction  $E_\alpha$  across the full context spectrum  $\alpha \in \{0.1, \dots, 0.9\}$  for all 38 TabArena classification datasets.

Table 1: Per-dataset effective baseline sample fraction  $E_\alpha$  and Mean Gain ( $E_\alpha - \alpha$ ) across 38 TabArena classification datasets for TabPFN (PFN) and TabFM (FM). Shaded green cells indicate effective samples ( $E_\alpha > \alpha$  or Mean Gain  $> 0$ ). Dashes (—) denote slices excluded by the teacher superiority filter (Section 4).

| Dataset                                    | 0.1  |      | 0.2  |      | 0.3  |      | 0.4  |      | 0.5  |      | 0.6  |      | 0.7  |      | 0.8  |      | 0.9  |      | Mean Gain |       |
|--------------------------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-----------|-------|
|                                            | PFN  | FM   | PFN  | FM   | PFN  | FM   | PFN  | FM   | PFN  | FM   | PFN  | FM   | PFN  | FM   | PFN  | FM   | PFN  | FM   | PFN       | FM    |
| APSFaure                                   | 0.05 | 0.17 | 0.22 | 0.24 | 0.28 | 0.30 | 0.38 | 0.65 | 0.50 | 0.25 | 0.59 | 0.70 | 0.70 | 0.70 | 0.69 | 0.72 | 0.91 | 0.93 | —         | -0.02 |
| Amazon_employee_access                     | 0.12 | 0.13 | 0.20 | 0.21 | 0.30 | 0.31 | 0.41 | 0.42 | 0.52 | 0.53 | 0.62 | 0.62 | 0.72 | 0.72 | 0.83 | 0.83 | 0.93 | 0.93 | 0.93      | +0.02 |
| Bank_Customer_Churn                        | 0.29 | 0.20 | 0.40 | 0.29 | 0.58 | 0.45 | 0.68 | 0.61 | 0.74 | 0.91 | 0.80 | 0.84 | 0.85 | 0.85 | 0.94 | 0.89 | 0.94 | 0.92 | 0.95      | +0.18 |
| Bioreponse                                 | 0.12 | 0.13 | 0.19 | 0.24 | 0.26 | 0.35 | 0.36 | 0.44 | 0.46 | 0.50 | 0.55 | 0.58 | 0.64 | 0.69 | 0.76 | 0.81 | 0.91 | 0.91 | 0.91      | -0.03 |
| Diabetes130US                              | 0.16 | 0.17 | 0.25 | 0.34 | 0.34 | 0.47 | 0.42 | 0.70 | 0.49 | 0.70 | 0.58 | 0.74 | 0.72 | 0.78 | 0.82 | 0.89 | 0.93 | 0.93 | 0.93      | +0.02 |
| E-CommerceShippingData                     | 0.91 | 0.10 | 0.46 | 0.10 | 1.00 | 0.10 | 0.98 | 0.10 | 1.00 | —    | 1.00 | —    | 1.00 | 0.85 | 0.96 | 0.91 | 0.94 | —    | —         | +0.42 |
| Fitness Club                               | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | —    | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90      | +0.41 |
| GiveMeSomeCredit                           | 0.21 | 0.19 | 0.30 | 0.31 | 0.40 | 0.43 | 0.46 | 0.46 | 0.51 | 0.55 | 0.64 | 0.55 | 0.69 | 0.62 | 0.78 | 0.74 | 0.86 | 0.88 | 0.88      | +0.03 |
| HR_Analytics_Job_Change_of_Data_Scientists | 0.40 | 0.21 | 0.52 | 0.36 | 0.71 | 0.40 | 0.75 | 0.38 | 0.77 | 0.62 | 0.78 | 0.67 | 0.77 | 0.75 | 0.89 | 0.81 | 0.95 | 0.97 | 0.97      | +0.23 |
| Is-this-a-good-customer                    | 0.47 | 0.95 | 0.89 | 0.93 | 0.97 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 0.99 | 1.00 | 0.98 | 1.00 | 0.97 | 0.99 | 1.00 | 0.96 | 0.96      | +0.42 |
| MIC                                        | 0.00 | 0.04 | 0.13 | 0.16 | 0.28 | 0.23 | 0.30 | 0.26 | 0.37 | 0.35 | 0.48 | 0.41 | 0.49 | 0.44 | 0.79 | 0.43 | 0.86 | 0.65 | —         | -0.17 |
| Marketing_Campaign                         | 0.69 | 0.69 | 0.79 | 0.62 | 0.80 | 0.84 | 0.73 | 0.80 | 0.68 | 0.80 | 0.78 | 0.80 | 0.84 | 0.81 | 0.88 | 0.85 | 1.00 | 0.92 | 0.92      | +0.30 |
| NATICSdroid                                | 0.22 | 0.23 | 0.35 | 0.34 | 0.44 | 0.41 | 0.53 | 0.51 | 0.58 | 0.58 | 0.64 | 0.63 | 0.82 | 0.75 | 0.84 | 0.82 | 0.91 | 0.92 | 0.92      | +0.09 |
| SDSS17                                     | 0.06 | 0.13 | 0.16 | 0.20 | 0.23 | 0.28 | 0.30 | 0.39 | 0.42 | 0.48 | 0.60 | 0.56 | 0.69 | 0.71 | 0.80 | 0.79 | 0.85 | 0.91 | —         | -0.04 |
| anneal                                     | 0.12 | 0.18 | 0.27 | 0.35 | 0.44 | 0.41 | 0.53 | 0.52 | 0.58 | 0.58 | 0.74 | 0.77 | 0.84 | 0.84 | 0.90 | 0.91 | 0.96 | 0.96 | 0.96      | +0.10 |
| bank-marketing                             | 0.44 | 0.35 | 0.67 | 0.39 | 0.69 | 0.45 | 0.75 | 0.60 | 0.78 | 0.63 | 0.83 | 0.66 | 0.89 | 0.80 | 0.95 | 0.89 | 0.96 | 0.95 | 0.95      | +0.27 |
| blood-transfusion-service-center           | 0.54 | 0.49 | 0.48 | 0.39 | 0.61 | 0.52 | 0.66 | 0.59 | 0.68 | 0.74 | 0.70 | 0.78 | 0.84 | 0.76 | —    | —    | 0.85 | 0.85 | 0.85      | +0.18 |
| churn                                      | 0.21 | 0.19 | 0.29 | 0.19 | 0.56 | 0.27 | 0.62 | 0.90 | 0.69 | —    | 0.81 | 0.45 | 0.91 | 0.90 | 0.91 | 0.90 | 0.93 | 0.90 | 0.90      | +0.16 |
| coil2000_insurance_policies                | 0.42 | 0.28 | 0.47 | 0.36 | 0.66 | 0.43 | 0.71 | 0.60 | 0.74 | 0.67 | 0.83 | 0.70 | 0.79 | 0.71 | 0.87 | 0.72 | 0.96 | —    | —         | +0.21 |
| credit-g                                   | 0.58 | 0.57 | 0.66 | 0.57 | 0.76 | 0.63 | 0.81 | 0.78 | 0.79 | 0.80 | 0.82 | 0.94 | 0.89 | 0.94 | 0.89 | 1.00 | 0.95 | 1.00 | 1.00      | +0.30 |
| credit_card_clients_default                | 0.13 | 0.11 | 0.18 | 0.17 | 0.28 | 0.28 | 0.38 | 0.39 | 0.48 | 0.47 | 0.59 | 0.55 | 0.68 | 0.65 | 0.77 | 0.70 | 0.92 | 0.77 | —         | -0.01 |
| customer_satisfaction_in_airline           | 0.12 | 0.08 | 0.22 | 0.19 | 0.32 | 0.28 | 0.41 | 0.38 | 0.51 | 0.45 | 0.61 | 0.57 | 0.70 | 0.66 | 0.80 | 0.78 | 0.89 | 0.88 | 0.88      | +0.01 |
| diabetes                                   | 0.77 | 1.00 | 0.79 | 0.80 | 1.00 | 0.84 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.85 | 0.95 | 0.96 | —    | —    | 0.92 | —    | —         | +0.53 |
| hazelnut-spread-contaminant-detection      | 0.30 | 0.20 | 0.38 | 0.34 | 0.46 | 0.40 | 0.55 | 0.51 | 0.60 | 0.58 | 0.72 | 0.66 | 0.77 | 0.74 | 0.88 | 0.81 | 0.94 | 0.93 | 0.93      | +0.08 |
| heloc                                      | 0.05 | 0.26 | 0.14 | 0.29 | 0.25 | 0.35 | 0.39 | 0.44 | 0.56 | 0.50 | 0.59 | 0.54 | 0.71 | 0.60 | 0.97 | 0.76 | 0.97 | 0.69 | 0.69      | +0.01 |
| hiva_agnostic                              | —    | 0.90 | —    | 0.90 | —    | 0.90 | —    | 0.90 | —    | 0.90 | —    | —    | —    | 0.90 | —    | 0.90 | 0.90 | 0.90 | 0.90      | -0.83 |
| in_vehicle_coupon_recommendation           | 0.21 | 0.15 | 0.31 | 0.25 | 0.37 | 0.32 | 0.45 | 0.43 | 0.56 | 0.52 | 0.65 | 0.63 | 0.73 | 0.73 | 0.82 | 0.81 | 0.91 | 0.90 | 0.90      | +0.06 |
| iml                                        | 0.22 | 0.17 | 0.26 | 0.24 | 0.32 | 0.30 | 0.43 | 0.42 | 0.55 | 0.50 | 0.63 | 0.55 | 0.73 | 0.68 | 0.81 | 0.79 | 0.91 | 0.90 | 0.90      | +0.04 |
| kddcup09_appetency                         | 0.15 | 0.16 | 0.27 | 0.22 | 0.30 | 0.24 | 0.37 | 0.38 | 0.38 | 0.38 | 0.63 | 0.63 | 0.66 | 0.71 | 0.70 | 0.76 | 0.91 | 0.87 | 0.61      | -0.00 |
| maternal_health_risk                       | 0.12 | 0.14 | 0.23 | 0.20 | 0.33 | 0.26 | 0.42 | 0.37 | 0.59 | 0.46 | 0.62 | 0.58 | 0.72 | 0.69 | 0.80 | 0.78 | 0.91 | 0.89 | 0.89      | +0.03 |
| online_shoppers_intention                  | 0.42 | 0.35 | 0.50 | 0.47 | 0.58 | 0.47 | 0.72 | 0.56 | 0.79 | 0.58 | 0.81 | 0.73 | 0.81 | 0.82 | 0.89 | 0.85 | 0.92 | 0.96 | 0.96      | +0.21 |
| polish_companies_bankruptcy                | 0.12 | 0.10 | 0.23 | 0.19 | 0.29 | 0.28 | 0.38 | 0.37 | 0.47 | 0.49 | 0.57 | 0.56 | 0.71 | 0.68 | 0.68 | 0.85 | 0.87 | 0.87 | 0.87      | -0.02 |
| qsar-biodeg                                | 0.50 | 0.59 | 0.62 | 0.71 | 0.67 | 0.72 | 0.72 | 0.84 | 0.77 | 0.88 | 0.82 | 0.90 | 0.88 | 0.91 | 0.91 | 0.94 | 1.00 | 0.99 | 0.99      | +0.27 |
| seismic-bumps                              | 0.72 | 0.71 | 0.74 | 0.51 | 0.71 | 0.73 | 0.71 | 0.79 | 0.75 | 0.83 | 0.77 | 0.74 | 0.79 | 0.88 | 0.99 | 0.85 | 0.92 | 0.79 | 0.79      | +0.26 |
| splice                                     | 0.24 | 0.18 | 0.48 | 0.30 | 0.55 | 0.48 | 0.64 | 0.69 | 0.68 | 0.75 | 0.76 | 0.75 | 0.83 | 0.78 | 0.91 | 0.92 | 0.95 | 0.95 | 0.95      | +0.17 |
| students_dropout_and_academic_success      | 0.29 | 0.30 | 0.41 | 0.35 | 0.49 | 0.42 | 0.56 | 0.50 | 0.61 | 0.63 | 0.66 | 0.66 | 0.73 | 0.78 | 0.83 | 0.83 | 0.89 | 0.91 | 0.91      | +0.10 |
| taiwanese_bankruptcy_prediction            | 0.31 | 0.26 | 0.41 | 0.35 | 0.39 | 0.39 | 0.46 | 0.50 | 0.55 | 0.62 | 0.66 | 0.66 | 0.73 | 0.78 | 0.83 | 0.83 | 0.89 | 0.91 | 0.91      | +0.04 |
| website_phishing                           | 0.29 | 0.48 | 0.43 | 0.59 | 0.53 | 0.76 | 0.63 | 0.89 | 0.75 | 0.92 | 0.86 | 0.93 | 0.93 | 0.97 | 0.95 | 0.98 | 0.97 | 0.99 | 0.99      | +0.33 |
| Average                                    | 0.32 | 0.33 | 0.41 | 0.38 | 0.51 | 0.46 | 0.58 | 0.58 | 0.64 | 0.64 | 0.71 | 0.68 | 0.79 | 0.77 | 0.86 | 0.84 | 0.90 | 0.89 | 0.89      | +0.12 |