Tabular Data War: Can Deep Learning Conquer the Last Bastion of GBDT?

2026. 01. 23

Data Mining & Quality Analytics Lab.

이 용 우

발표자 소개

❖ 이 용 우 (Yongwoo Lee)

- 고려대학교 일반대학원 산업경영공학과 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석사 과정 2학기차 (2025. 09 ~)

❖ Research Interest

- Tabular Data Deep Learning
- Domain Adaptation/Generalization

❖ Contact

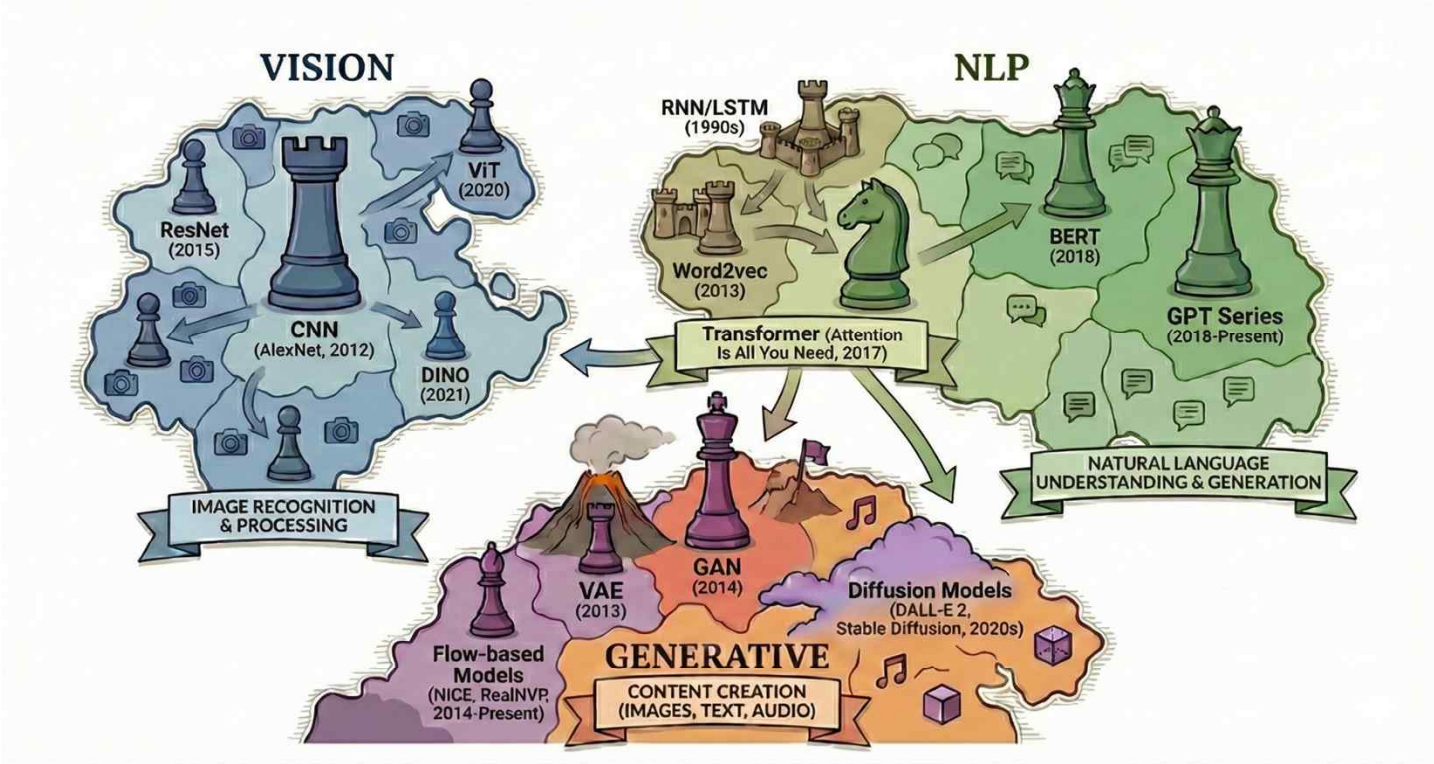
- ywlee1@korea.ac.kr

Contents

1. Introduction

Introduction

The Golden Age of Deep Learning: Conquering Unstructured Data



"The World is conquered by Deep Learning"

AI Generated

Introduction

The Last Bastion: Why not Tabular?

2022

Why do tree-based models still outperform deep learning on typical tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
MLIA, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

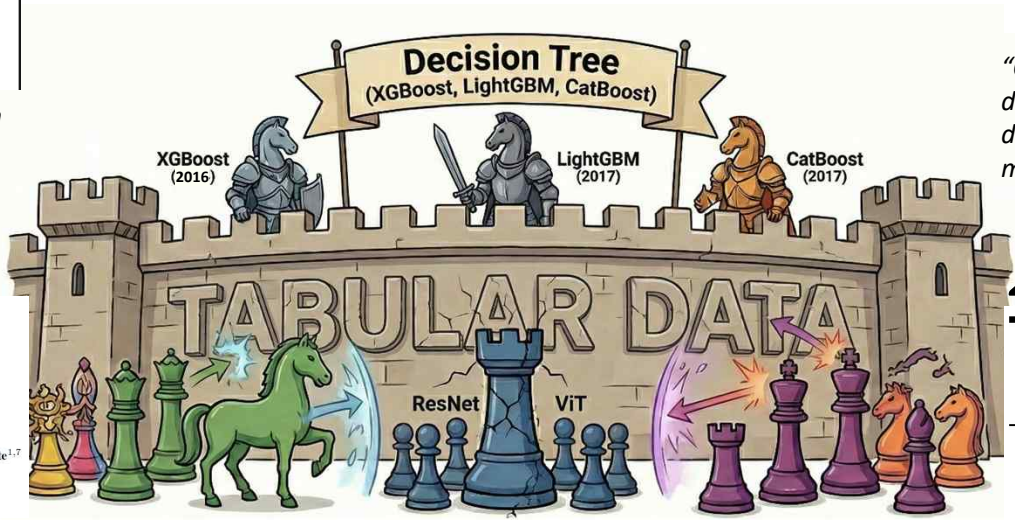
"Tree-based models remain state-of-the-art on medium-sized tabular data."

2023

When Do Neural Nets Outperform Boosted Trees on Tabular Data?

Duncan McElfresh[*1,2], Sujay Khandagale[3], Jonathan Valverde[4], Vishak Prasad C[5], Benjamin Feuer[6], Chinmay Hegde[6], Ganesh Ramakrishnan[5], Micah Goldblum[6], Colin White[1,7]
[1] Abacus.AI, [2] Stanford, [3] Pinterest, [4] University of Maryland,
[5] IIT Bombay, [6] New York University, [7] Caltech

"Despite recent advances in neural network (NN) architectures for tabular data, there is still an active debate over whether or not NNs generally outperform gradient-boosted decision trees (GBDTs) on tabular data"



2022

TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED

Ravid Shwartz-Ziv
ravid.ziv@intel.com
IT AI Group, Intel

Amitai Armon
amitai.armon@intel.com
IT AI Group, Intel

November 24, 2021

"Our study shows that XGBoost outperforms these deep models across the datasets, including the datasets used in the papers that proposed the deep models."

2024

A Data-Centric Perspective on Evaluating Machine Learning Models for Tabular Data

Andrej Tschalzev*
University of Mannheim

Sascha Marton
University of Mannheim

Stefan Lüdtke
University of Rostock

Christian Bartelt
University of Mannheim

Heiner Stuckenschmidt
University of Mannheim

"Despite these efforts, Gradient Boosted Decision Trees (GBDTs) remain the state-of-the-art."

Grinsztajn et al., "Why do tree-based models still outperform deep learning...?", NeurIPS 2022 (Abstract).
Shwartz-Ziv & Armon, "Tabular Data: Deep Learning is Not All You Need", Information Fusion 2022 (Abstract).
Jiang et al., "A Data-Centric Perspective on Evaluating Machine Learning Models...", ArXiv 2024 (Abstract).

The Tabular Data War (2022-2025)

Background: Cleaning up the Hype

NEURAL OBLIVIOUS DECISION ENSEMBLES FOR DEEP LEARNING ON TABULAR DATA

TabNet: Attentive Interpretable Tabular Learning

"We have beaten GBDT"

Sercan Ö. Arık, Tomas Pfister
Google Cloud AI
Sunnyvale, CA
soarik@google.com, tpfister@google.com

Revisiting Deep Learning Models for Tabular Data

"No, you haven't"

Yury Gorishniy† … Artem Babenko†♣

† Yandex, Russia
‡ Moscow Institute of Physics and Technology, Russia
♣ National Research University Higher School of Economics, Russia

2022 NeurIPS, 인용수 : 2498회 (2026/1/23)

Why do tree-based models still outperform deep learning on typical tabular data?

Edouard Oyallon
MLIA, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

leo.grinsztajn@inria.fr

Issues:

- **Unequal Tuning**
- **No Benchmark**
- **Unreplicability**

While deep learning has enabled tremendous progress on text and image datasets, its superiority on tabular data is not clear. We contribute extensive benchmarks of standard and novel deep learning methods as well as tree-based models such as XGBoost and Random Forests, across a large number of datasets and hyperparameter combinations. We define a standard set of 45 datasets from varied domains with clear characteristics of tabular data and a benchmarking methodology accounting for both fitting models and finding good hyperparameters. Results show that tree-based models remain state-of-the-art on medium-sized data (10K samples) even without accounting for their superior speed. To understand this gap, we conduct an empirical investigation into the differing inductive biases of tree-based models and neural networks. This leads to a series of challenges which should guide researchers aiming to build tabular-specific neural network: 1. be robust to uninformative features, 2. preserve the orientation of the data, and 3. be able to easily learn irregular functions. To stimulate research on tabular architectures, we contribute a standard benchmark and raw data for baselines: every point of a 20 000 compute hours hyperparameter search for each learner.

Data Mining Quality Analytics

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems, 35

The Tabular Data War (2022-2025)

The Benchmark: 45 OpenML Datasets

❖ 데이터셋 선정 기준 (OpenML에서 **총 45개 데이터셋 엄선**)

- **이질적 컬럼 (Heterogeneous columns)**
- **낮은 차원 (Not high dimensional):** 샘플 수 대비 차원 비율 **1/10 미만**, 전체 차원 수는 **500 미만**
- **문서화 여부 (Undocumented datasets):** 정보가 너무 부족한 데이터는 제외
- **I.I.D. 데이터 (I.I.D. data):** 독립 항등 분포를 따르는 데이터만 선택
- **실제 데이터 (Real-world data):** 인공(Artificial) 데이터 제외, 일부 **시뮬레이션 데이터**는 포함
- **적절한 크기 (Not too small):** 특성(feature) 4개 미만, 샘플 3,000개 미만 데이터 제외, 10,000개 이상은 학습 샘플을 10,000개 로 맞춤.
- **적절한 난이도 (Not too easy):** 단순 모델과 강력한 모델의 성능 차이가 **5% 미만**이면 제외
- **비결정론적 (Not deterministic):** 타겟이 데이터의 결정론적 함수인 경우 제외

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems, 35

The Tabular Data War (2022-2025)

The Fairness: Rigorous Hyperparameter Tuning

- **사용 모델 :**

DL	Tree Based
- **MLP** - **ResNet(Tabular version)** - **FT-Transformer** - **SAINT**	- **XGBoost** - **Random Forest** - **Gradient Boosting**

- **하이퍼 파라미터 최적화 프로토콜 :**
  - 각 데이터셋마다 **400회(Iterations)의 랜덤 서치 기회**를 부여.
  - Hyperopt-Sklearn 하이퍼파라미터 공간
  - Hyperopt-Sklearn에 해당 모델의 공간이 없거나, 원논문에서 특정 **모델에 대한 제안이** 있다면, 해당 논문의 권장 탐색 공간을 따름
  - 랜덤서치 순서를 15가지로 셔플(shuffle)하여 평균(순서에 따른 조기 **최적화 효과 방지**)

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems, 35

The Tabular Data War (2022-2025)

The Metric: Aggregation Strategy

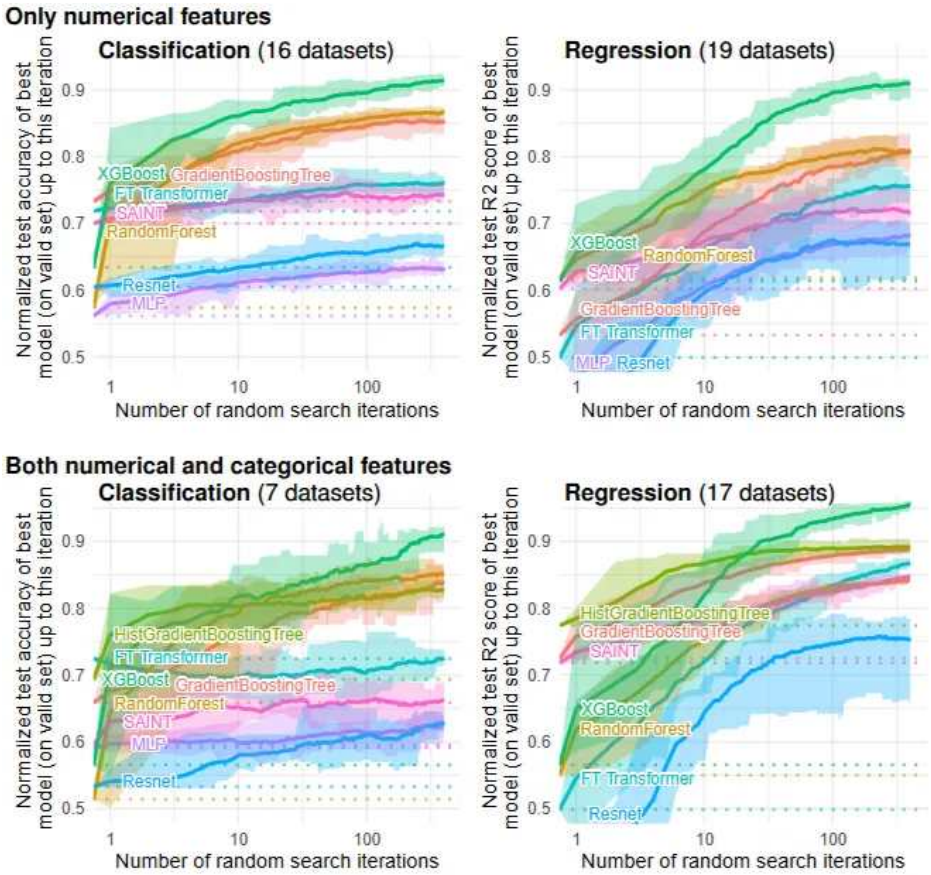
1. 평가 지표 (Metrics):
   - Classification: Test Accuracy (정확도)
   - Regression: R2 Score (설명력)

2. 통합 방법 (Aggregation Strategy):
   - ADTM(average distance to the minimum): 모든 점수를 **0(하위권) ~ 1(최고점)**로 변환하여 통합.
   - 하위권(0점) 기준 : 하위 10%(분류문제) / 하위 50%(회귀문제)
   - 0점의 기준을 '꼴찌 모델(Worst)'로 잡지 않음으로써 이상치 왜곡 방지.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems, 35

The Tabular Data War (2022-2025)

The Verdict: Trees Win on Medium Data

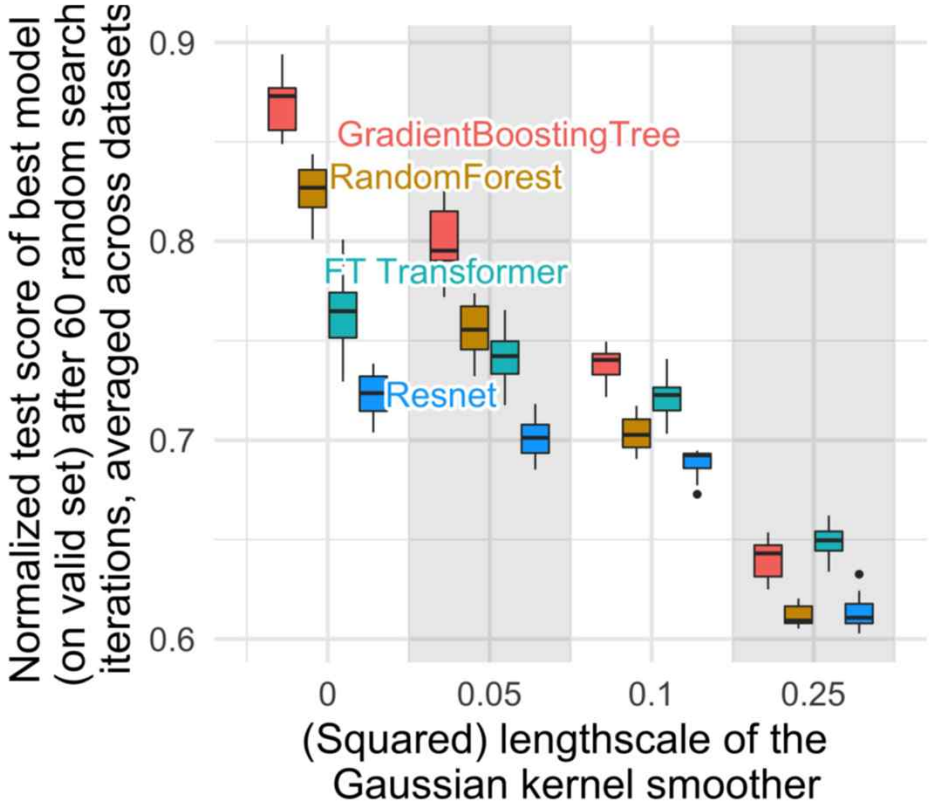


- **점선(Dotted lines) : 기본 하이퍼파라미터**에서 점수
- **값(Values) :** 특정 횟수만큼 랜덤 서치(튜닝)를 진행했을 때 찾은 최고 모델(Validation set 기준)의 테스트 점수.
- **신뢰 구간 :** 랜덤 서치 순서를 15번 섞어서(Shuffle) 평균을 낸 값이며, 리본(Ribbon) 영역은 최소/최대 점수 범위를 나타냄.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems, 35

The Tabular Data War (2022-2025)

Why DL Failed 1: overly smooth solution



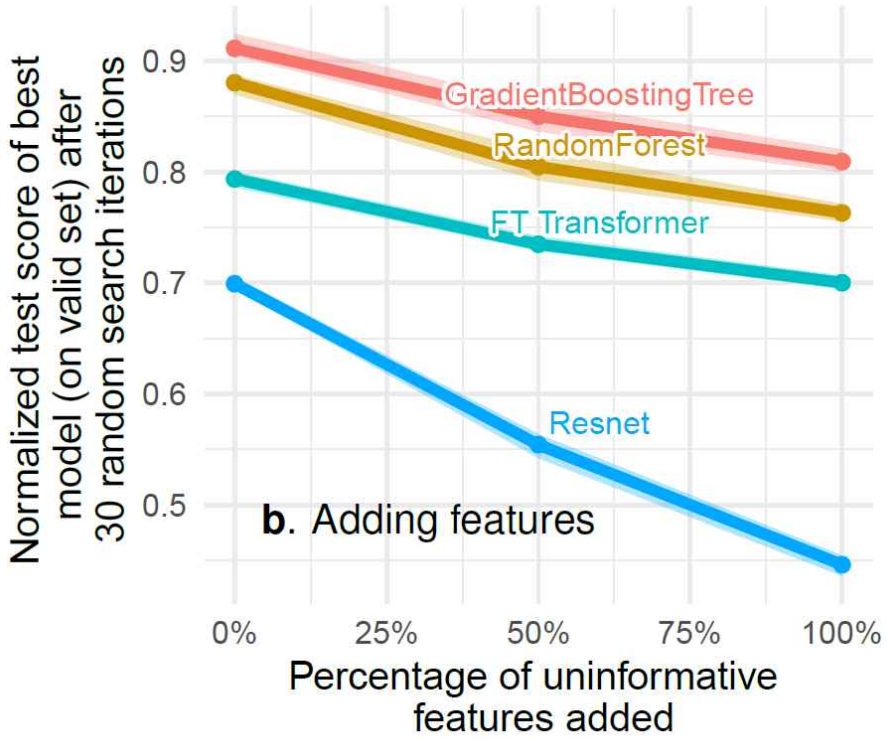
1. 실험 설계 (The Experiment):
   - 가설 : 정형 데이터의 타겟 함수는 종종 불연속적(Irregular)지만 딥러닝 모델은 **연속적이고 매끄러운 함수 근사를 유도하는 경향이 있다.**
   - 실험 조작: '학습' 데이터의 정답(Target) 함수를 Gaussian Kernel로 인위적으로 Smoothing '부드러운 데이터'로 변조함.
2. X축: Smoothing 정도 (0 ~ 0.25)
3. Y축: Normalized Test Score (모델 성능)

Data Mining Quality Analytics

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems, 35

The Tabular Data War (2022-2025)

Why DL Failed 2: Sensitivity to Uninformative Features



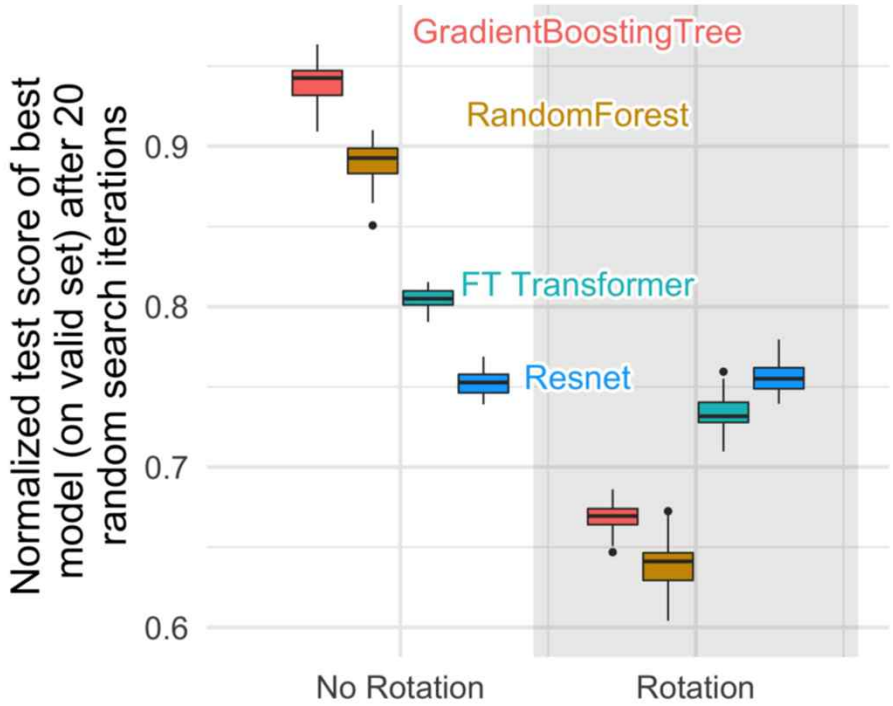
1. 실험 설계 (The Experiment):
   - 가설: ResNet과 같은 신경망 구조는 정보가 없는 불필요한(Uninformative) 특성에 취약하여, 이러한 특성이 늘어날수록 트리 기반 모델보다 성능이 더 크게 하락할 것이다.
   - 실험 조작: 기존 데이터셋에 타겟 변수(정답) 및 다른 특성들과 상관관계가 없는 무작위 노이즈(Standard Gaussians) 특성을 인위적으로 추가.
2. X축: 추가된 비정보적 특성의 비율(0~100%)
3. Y축: 최적 모델의 정규화된 **테스트 점수**

Data Mining Quality Analytics

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems, 35

The Tabular Data War (2022-2025)

Why DL Failed 3: Rotational Invariance



1. 실험 설계 (The Experiment):
   - 가설: 정형 데이터에서는 각 입력 특성 축이 고유한 의미를 가지므로, 입력 공간의 회전은 의미 있는 구조를 훼손하게 된다. 따라서 축에 민감한 트리 기반 모델은 회전에 의해 성능이 크게 저하되며, 회전 불변적인 성질을 가진 신경망 모델은 상대적으로 영향을 덜 받을 것이다.
   - 조작: 무작위 회전 행렬(Random Rotation Matrix) 적용
2. X축: No Rotation(원본 데이터) / Rotation(회전된 데이터)
3. Y축: 최적 모델의 정규화된 테스트 점수

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems, 35

The Tabular Data War (2022-2025)

A New Front Opens: TabArena

1. Problematic data :

- Outdated, problematic license, data leakage
- 현실(Real-world task)을 반영하지 못함.

2. Baselines Stuck in Time:

- 벤치마크가 한 번 발표되면 업데이트되지 않음(**Static**).
- 최신 모델(SOTA)이 나왔는데도, 여전히 몇 년 전 모델하고만 비교함.

3. Unreliable Evaluation ：

- 기존 평가 프로토콜에 대한 회의론(Skepticism) 증가.
- 후속 연구에서 결함이 발견되어도 수정되지 않음.

2025 NeurIPS, 인용수 : 29회 (2026/1/23)

TabArena: A Living Benchmark for Machine Learning on Tabular Data

Nick Erickson[1] Lennart Purucker[2] Andrej Tschalzev[3] David Holzmüller[4,5,6]
Prateek Mutalik Desai[1] David Salinas[8,2] Frank Hutter[7,8,2]
[1]Amazon Web Services [2]University of Freiburg [3]University of Mannheim [4]INRIA Paris
[5]Ecole Normale Supérieure [6]PSL Research University [7]Prior Labs [8]ELLIS Institute Tübingen
mail@tabarena.ai

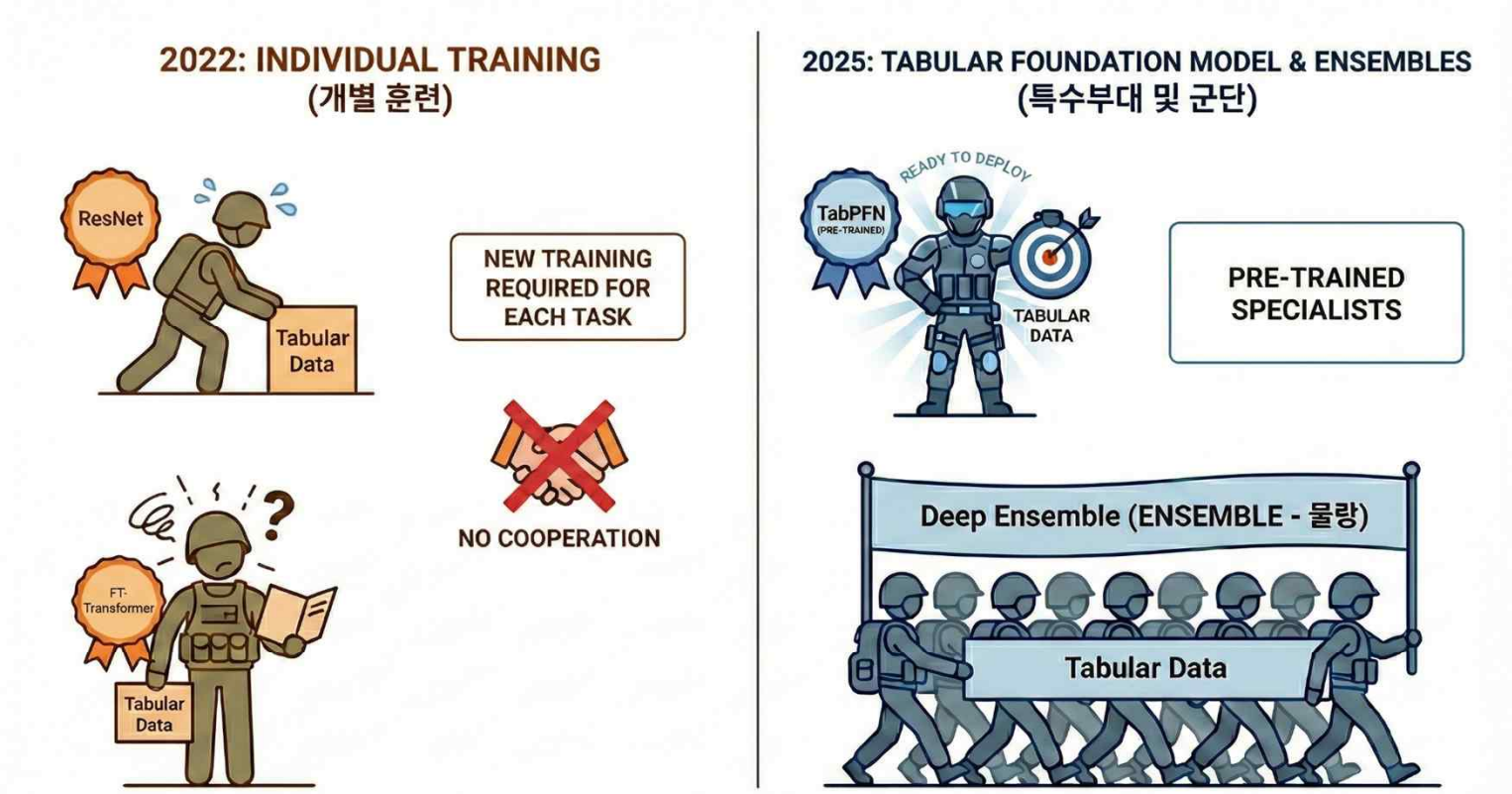
Abstract

With the growing popularity of deep learning and foundation models for tabular data, the need for standardized and reliable benchmarks is higher than ever. However, current benchmarks are static. Their design is not updated even if flaws are discovered, model versions are updated, or new models are released. To address this, we introduce TabArena, the first continuously maintained living tabular benchmarking system. To launch TabArena, we manually curate a representative collection of datasets and well-implemented models, conduct a large-scale benchmarking study to initialize a public leaderboard, and assemble a team of experienced maintainers. Our results highlight the influence of validation method and ensembling of hyperparameter configurations to benchmark models at their full potential. While gradient-boosted trees are still strong contenders on practical tabular datasets, we observe that deep learning methods have caught up under larger time budgets with ensembling. At the same time, foundation models excel on smaller datasets. Finally, we show that ensembles across models advance the state-of-the-art in tabular machine learning. We observe that some deep learning models are overrepresented in cross-model ensembles due to validation set overfitting, and we encourage model developers to address this issue. We launch TabArena with a public leaderboard, reproducible code, and maintenance protocols to create a living benchmark available at https://tabarena.ai.

Erickson, N., et al. (2025). TabArena: A Living Benchmark for Machine Learning on Tabular Data. Advances in Neural Information Processing Systems, 38.

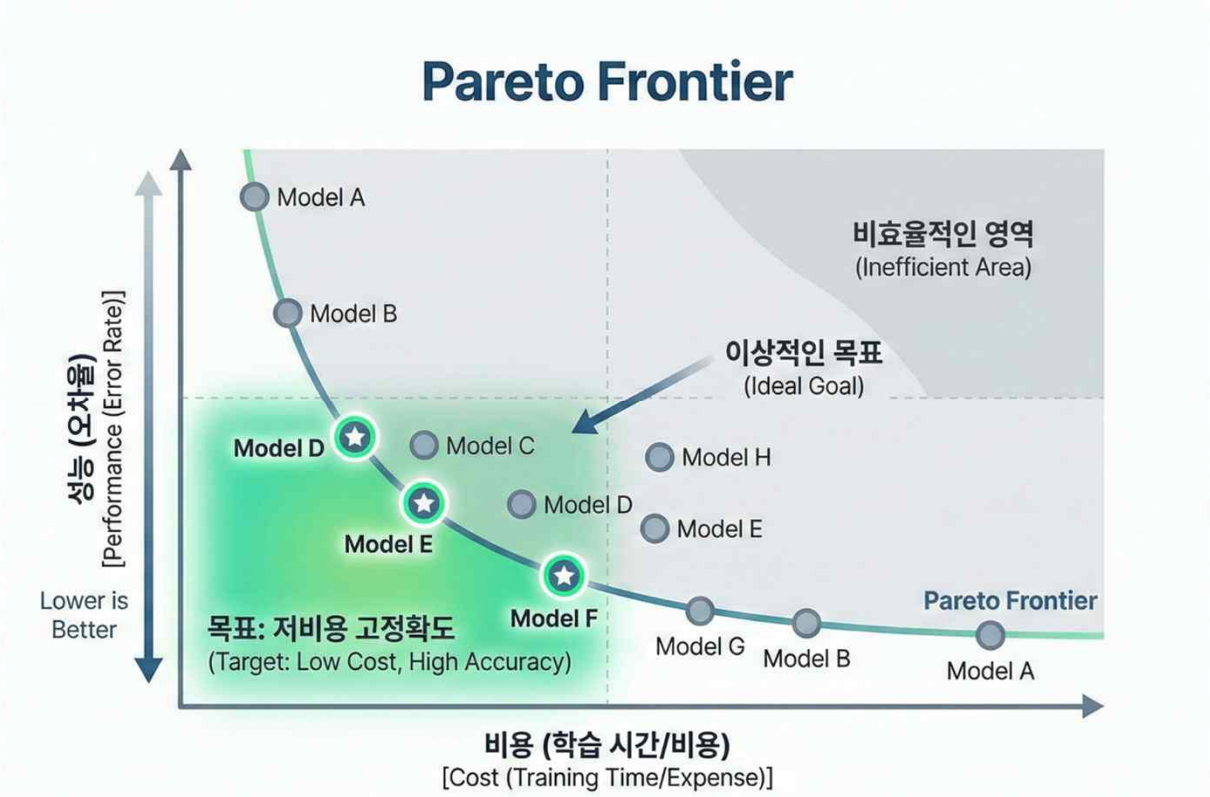
The Tabular Data War (2022-2025)

What is different from 2022 (1/2) : New Weapons - Pre-trained specialist & Ensembles



The Tabular Data War (2022-2025)

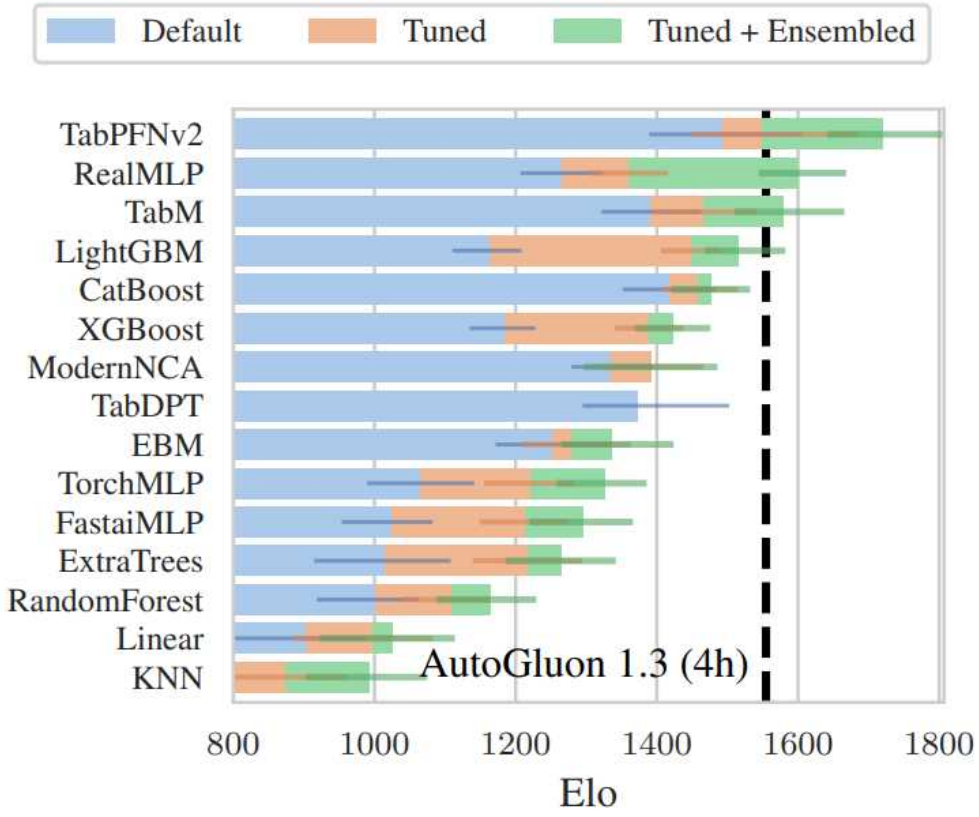
What is different from 2022 (2/2) : From 'Equal Iterations' to 'Equal Time'



1. 2022 (Scientific Fairness):
   - 기준: Iteration (반복 횟수)
   - 맹점: 일반적으로 신경망이 트리 모델보다 학습이 더 느림, 현업에선 불공정함.
2. 2025 (Production Fairness):
   - 기준: Time Budget (제한 시간).
   - Pareto Frontier (시간 대비 성능) 평가 도입.

The Tabular Data War (2022-2025)

The Small Data(<10K, feature <500) Leaderboard : Enter the Foundation Model



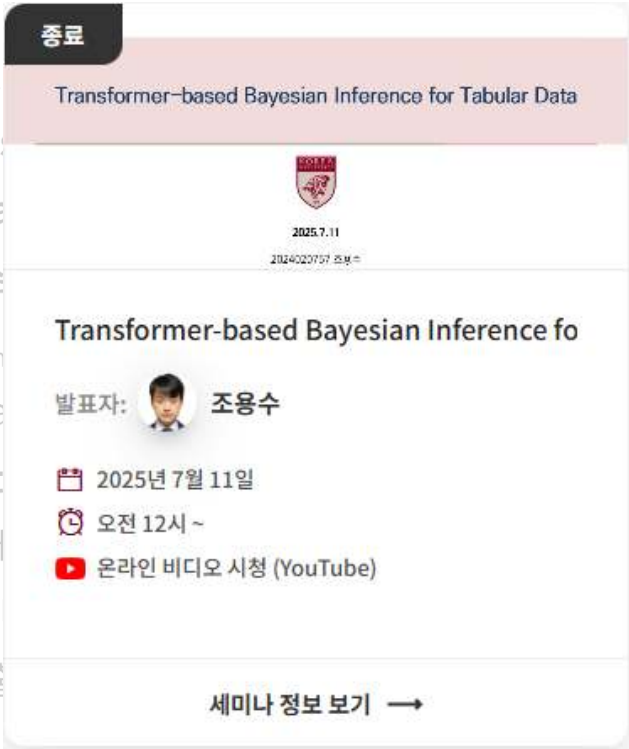
1. DataSet (tabPFNv2 사용 가능 데이터):
   - Small Data(데이터 1만 개 이하, 컬럼 500개 이하).
2. 평가지표 : Elo
   - 성능 우위(AUC, RMSE 등)를 **브래들리-테리(BT) 모델**로 통합 분석한 뒤, **랜덤 포레스트(1000점)를 기준**으로 환산하여 수치화한 점수
3. 결과 (Rank): TabPFNv2 (압도적, 2위와 격차 큼).
   - 튜닝이나 앙상블 없이, 기본값(파랑)만으로도 이미 최상위권.

Erickson, N., et al. (2025). TabArena: A Living Benchmark for Machine Learning on Tabular Data. Advances in Neural Information Processing Systems, 38.

The Tabular Data War (2022-2025)

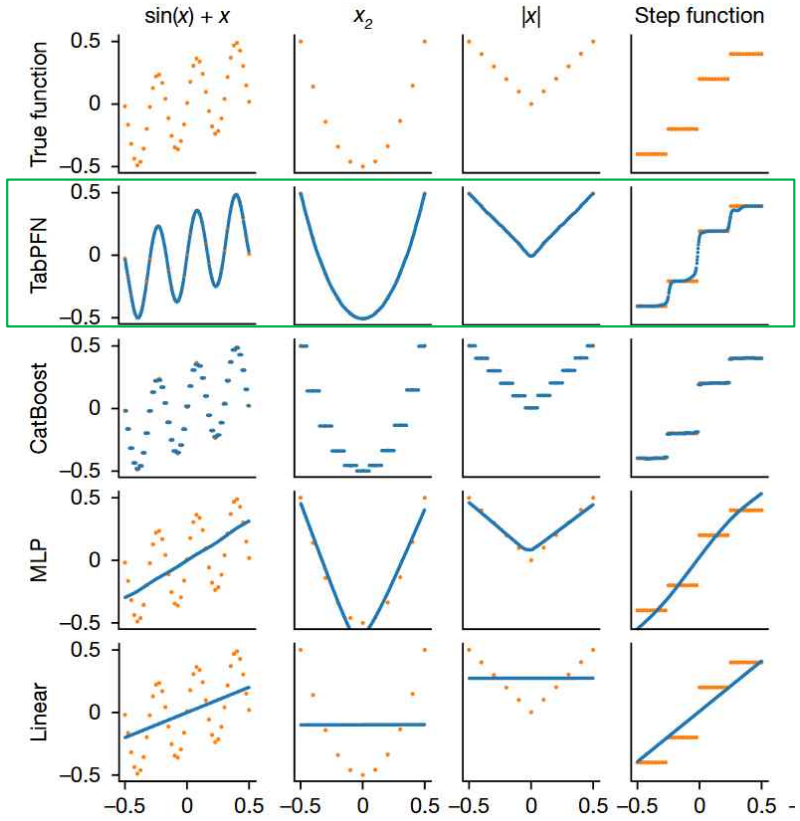
What is TabPFN?

- **1. 정의 (Definition):** In-Context 를 근사하는 트랜스포머 모델
- **2. 핵심 메커니즘 (Core Mechani**
  - **Training Phase (Meta-Lea**
    - **Input:** 수천만 개의 **가상**
    - **Objective:** $P(y|x,D_{train}$
    - **Source:** Structural Ca orks 등을 통해 생성된 다양한 분포.
  - **Inference Phase (In-Cont**
    - **Process:** 파라미터 업데
    - **Operation:** 전체 학습 나의 시퀀스로 입력 → **Single Forward P ass** → 예측값(y_{test}) 도출



The Tabular Data War (2022-2025)

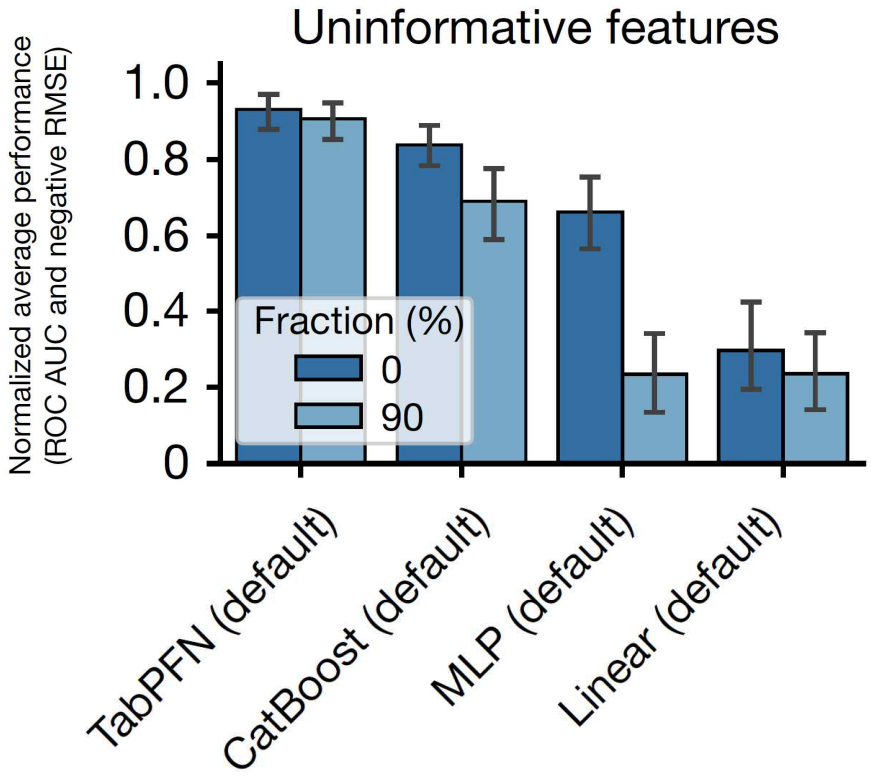
Revisiting the 2022 Bottleneck 1 : overly smooth solution



Hollmann, N., Müller, S., Purucker, L. et al. Accurate predictions on small data with a tabular foundation model. Nature 637, 319–326 (2025).

The Tabular Data War (2022-2025)

Revisiting the 2022 Bottleneck 2 : Uninformative Features

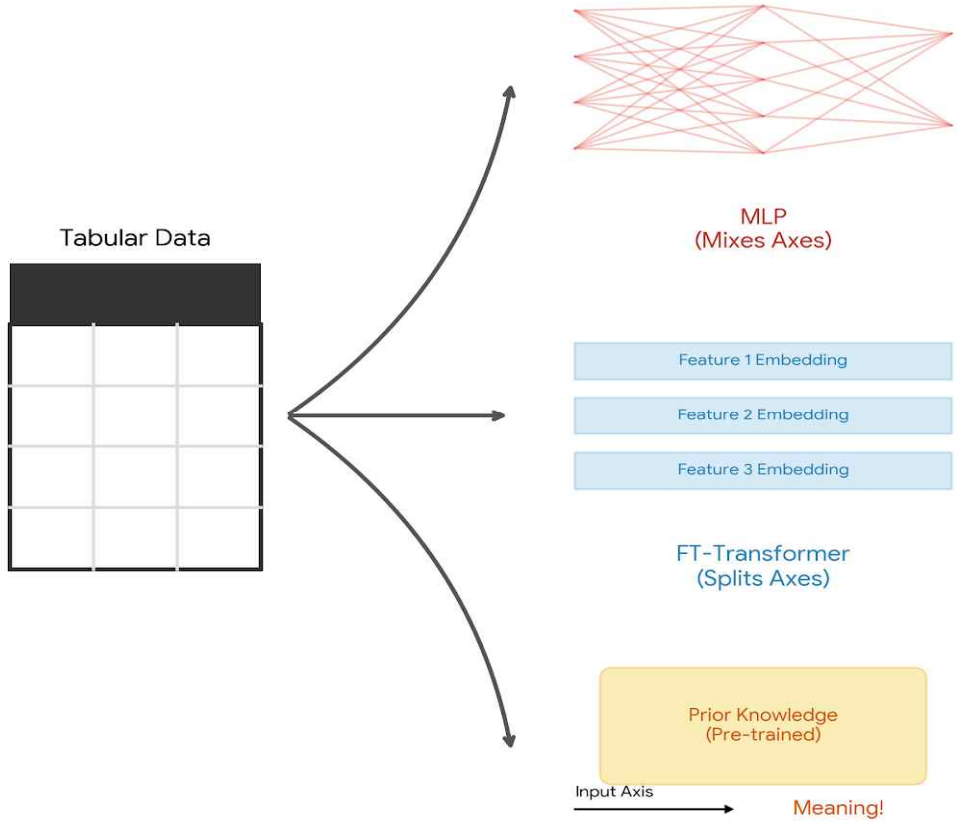


- 변화 요소 : 원본 데이터셋의 특성 D에 노이즈 특성을 9 x D개 추가(90%)
- 반복 횟수: 각 데이터셋 및 설정당 10회의 반복 실험 평균값
- 데이터셋 : AutoML Benchmark, OpenML-CTR23에서 선정된 분류 29개, 회귀 28개
- 평가 지표:
  - Classification: Normalized ROC AUC
  - Regression: Normalized Negative RMSE정규화
- Normalization 방식: min-max **normalization**

Hollmann, N., Müller, S., Purucker, L. et al. Accurate predictions on small data with a tabular foundation model. Nature 637, 319–326 (2025).

The Tabular Data War (2022-2025)

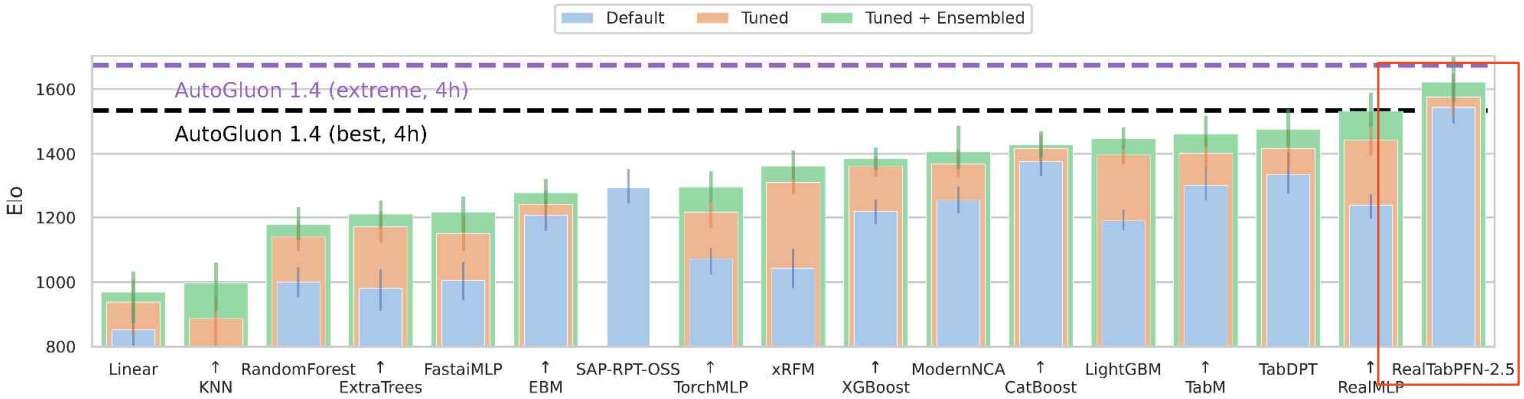
Revisiting the 2022 Bottleneck 3 : Rotational Invariance

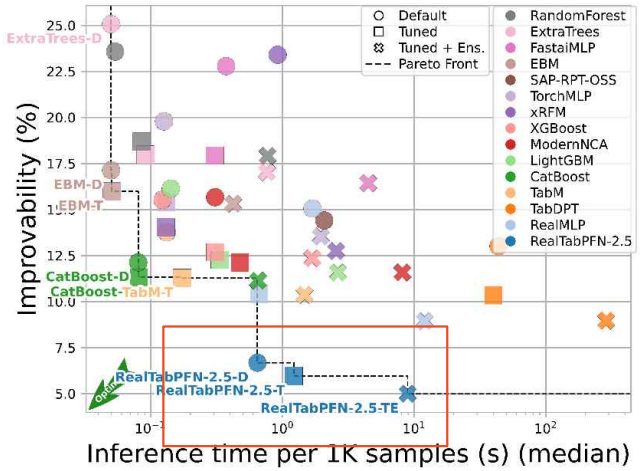


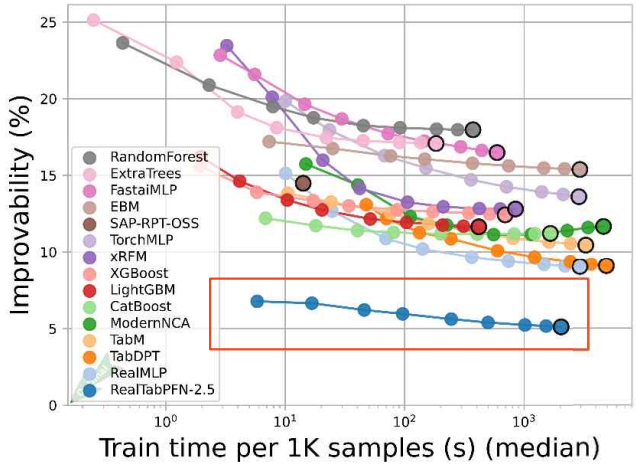
1. MLP : 축을 섞음
2. FT-Transformer, TabPFN : 축을 보존

The Tabular Data War (2022-2025)

Where Are We Now? — TabArena and TabPFN v2.5







3. Conclusion

Conclusions

Conclusion: A New Equilibrium in Tabular Learning

1. 딥러닝의 귀납적 편향 극복
    - The Barriers : Smoothness Bias, Uninformative Features, Rotational Invariance
    - 극복 메커니즘 : Synthetic Priors (TabPFN), Attention,
2. Tabular Foundation Model : Zero-shot, In-context Learning
3. Living Benchmark(tabArena) : 최적의 도구 실시간 확인

Thank You