Comparison of Machine/Deep learning Methods for Tabular Dataset

김경수

Korea University

Data **M**ining & **Q**uality **A**nalytics Lab.

22.09.30

KOREA UNIVERSITY Data Mining Quality Analytics

발표자 소재

- 김경수 (Kyungsoo Kim)
  - ✓ 고려대학교 산업경영공학과
  - ✓ Data Mining & Quality Analytics Lab. (김성범 교수님)
  - ✓ M.S Student (2022.03 ~ Present)
- Research Interest
  - ✓ Machine Learning & Deep Learning for tabular data
  - ✓ XAI
  - ✓ Anomaly Detection
- Contact
  - ✓ E-mail : kyungsoo@korea.ac.kr

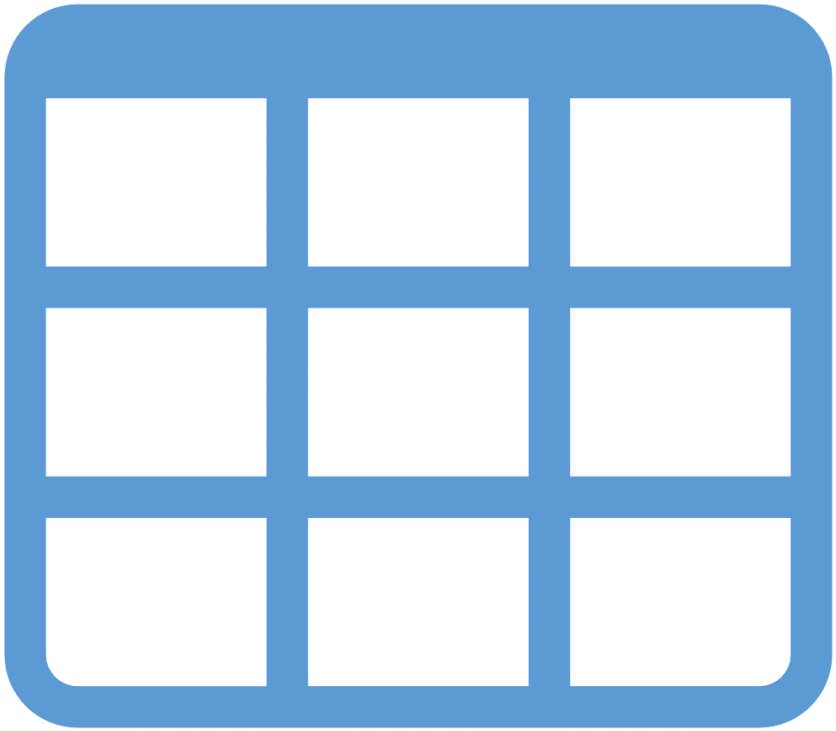
Contents

Introduction

Introduction

❖ Tabular Dataset ?

- 표 형태의 데이터로 주로 2차원 데이터
- 행렬로 구성된 데이터, 정형 데이터



https://matthewrenze.com/articles/working-with-tabular-data/

Introduction

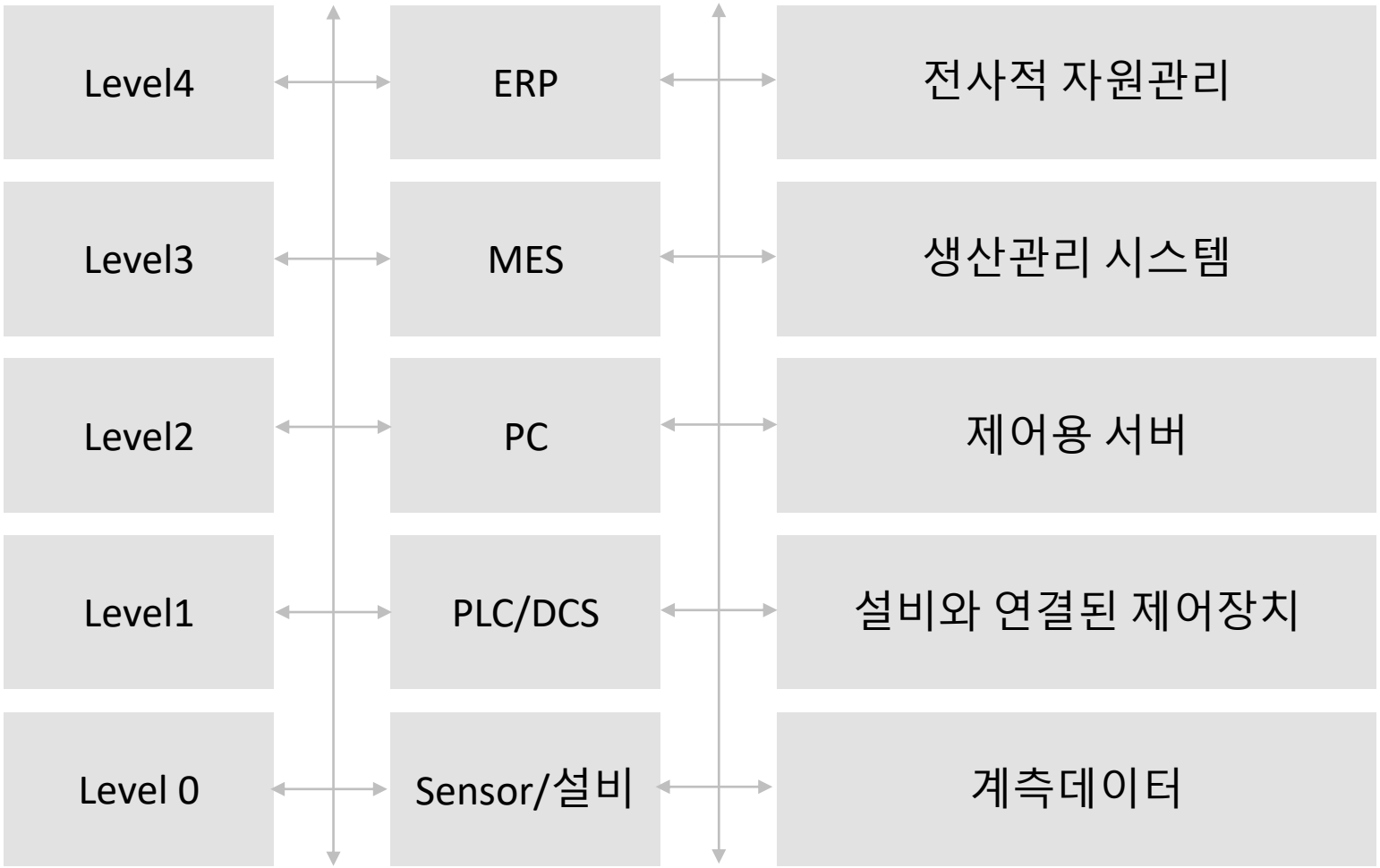
❖ Tabular Dataset 의 활용예시

- 제조업에서 데이터는 대부분 Table 형태의 데이터
- ERP: enterprise resource planning
- MES: Manufacturing Execution Systems
- P/C: Process Computer
- PLC: Programmable Logic Controller / DCS: Distributed Control System

Level4	ERP	전사적 자원관리
Level3	MES	생산관리 시스템
Level2	PC	제어용 서버
Level1	PLC/DCS	설비와 연결된 제어장치
Level0	Sensor	계측데이터

Introduction

❖ **Tabular Dataset 의 활용예시**

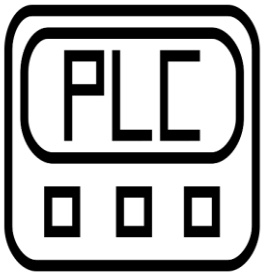


Introduction

❖ Tabular Dataset 의 활용예시

계측

시간	value
0:00:01	20
0:00:02	20
0:00:03	14
0:00:04	15
0:00:05	19



PLC

시간	value
0:00:01	20
0:00:02	20
0:00:03	14
0:00:04	15
0:00:05	19

P/C

시간	value
00~01	17
01~02	16
02~03	15
03~04	19
04~05	0

MES

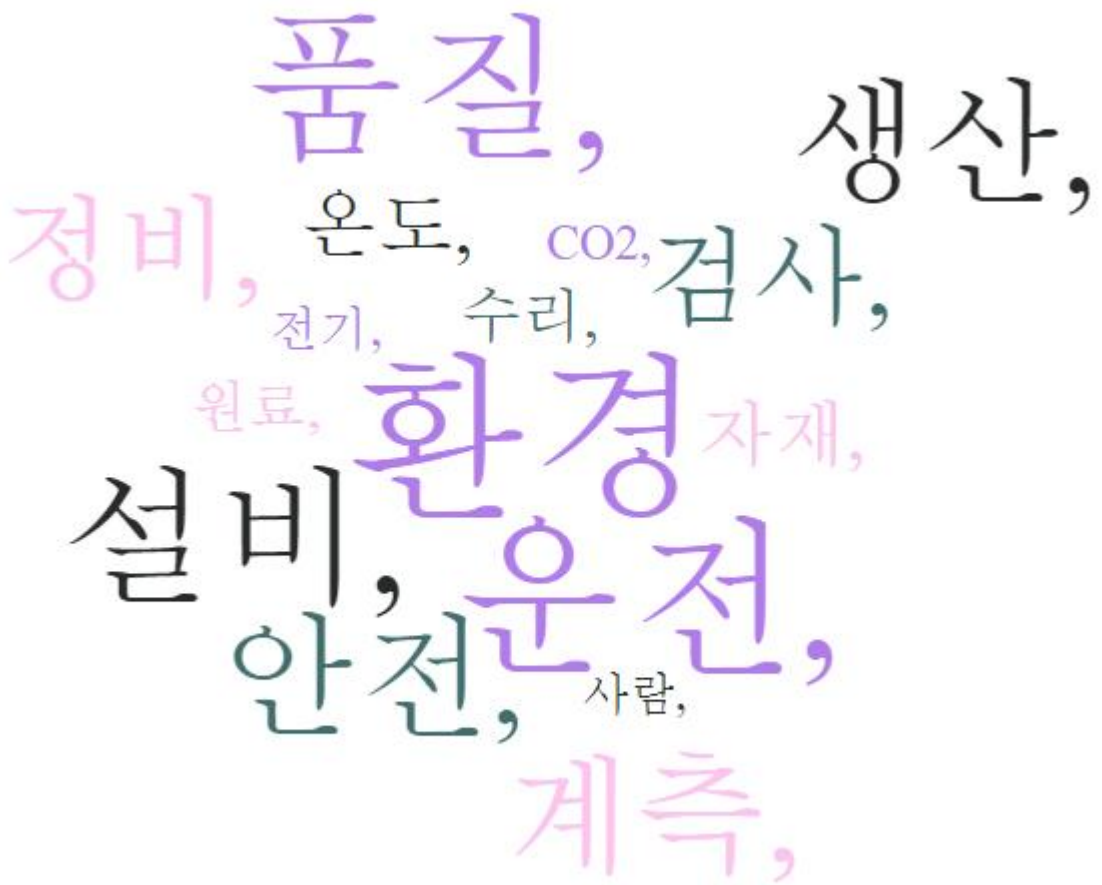
제품	불량
A	X
B	X
C	X
D	X
E	O

ERP

제품	원가
A	100
B	120
C	130
D	110
E	3000

Introduction

❖ Tabular Dataset를 위한 Machine/Deep Learning 필요성



Introduction

❖ Tabular Dataset를 위한 Machine/Deep Learning 필요성

Introduction

❖ Tabular Dataset를 위한 Machine/Deep Learning 필요성

설비	예지 정비
생산	수율 증가
검사	공정 생략

=

세미나 주제

Comparison of machine/deep learning methods for tabular dataset

Comparison of deep learning methods for tabular dataset

❖ Paper : Deep Neural Networks and Tabular Data : A Survey

- Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2022.06). (48회 인용)

Deep Neural Networks and Tabular Data: A Survey

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug,
Martin Pawelczyk and Gjergji Kasneci

***Abstract*—Heterogeneous tabular data are the most commonly used form of data and are essential for numerous critical and computationally demanding applications. On homogeneous data sets, deep neural networks have repeatedly shown excellent performance and have therefore been widely adopted. However, their adaptation to tabular data for inference or data generation tasks remains highly challenging. To facilitate further progress in the field, this work provides an overview of state-of-the-art deep learning methods for tabular data. We categorize these methods into three groups: data transformations, specialized architectures, and regularization models. For each of these groups, our work offers a comprehensive overview of the main approaches. Moreover, we discuss deep learning approaches for generating tabular data, and we also provide an overview over strategies for explaining deep models on tabular data. Thus, our first contribution is to address the main research streams and existing methodologies in the mentioned areas, while highlighting relevant challenges and open research questions. Our second contribution is to provide an empirical comparison of traditional machine learning methods with eleven deep learning approaches across five popular real-world tabular data sets of different sizes and with different learning objectives. Our results, which we have made publicly available as competitive benchmarks, indicate that algorithms based on gradient-boosted tree ensembles still mostly outperform deep learning models on supervised learning tasks, suggesting that the research progress on competitive deep learning models for tabular data is stagnating. To the best of our knowledge, this is the first in-depth overview of deep learning approaches for tabular data; as such, this work can serve as a valuable starting point to guide researchers and practitioners interested in deep learning with tabular data.**

***Index Terms*—Deep neural networks, Tabular data, Heterogeneous data, Discrete data, Tabular data generation, Probabilistic modeling, Interpretability, Benchmark, Survey**

I. Introduction

Ever-increasing computational resources and the availability of large, labelled data sets have accelerated the success of deep neural networks [1], [2]. In particular, architectures based on convolutions, recurrent mechanisms [3], or transformers [4] have led to unprecedented performance in a multitude of domains. Although deep learning methods perform outstandingly well for classification or data generation tasks on homogeneous data (e.g., image, audio, and text data), tabular data still pose a challenge to deep learning models [5]–[8]. Tabular data – in contrast to image or language data – are heterogeneous, leading to dense numerical and sparse categorical features. Furthermore, the correlation among the features is weaker than the one introduced through spatial or semantic relationships in image or speech data. Hence, it is necessary to discover and exploit relations without relying on spatial information [9]. Therefore, Kadra et al. called tabular data sets the last *"unconquered castle"* for deep neural network models [10].

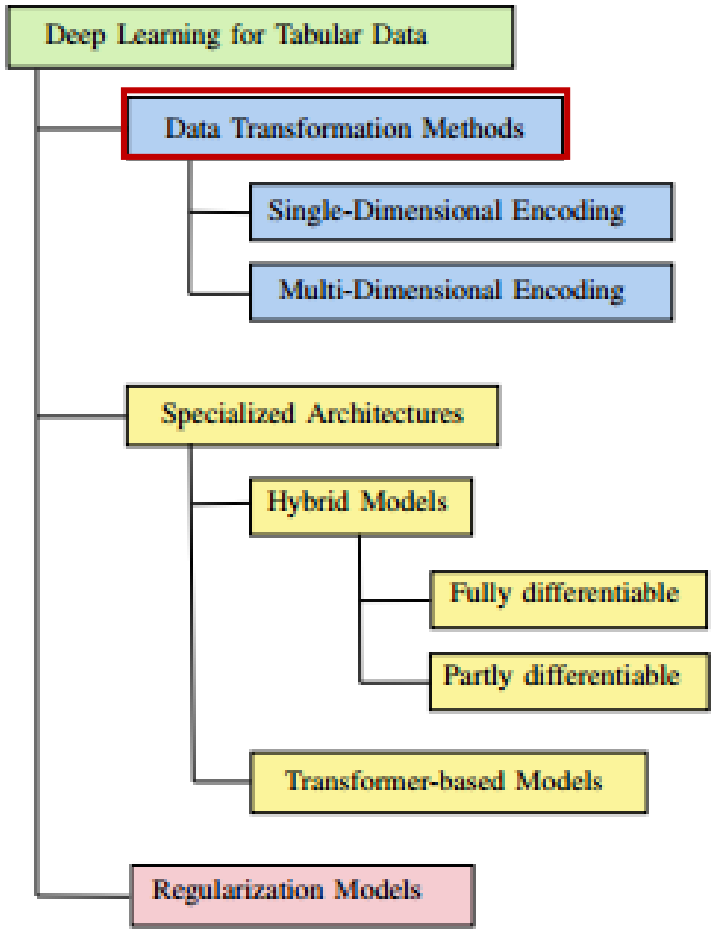
Heterogeneous data are the most commonly used form of data [7], and it is ubiquitous in many crucial applications, such as medical diagnosis based on patient history [11]–[13], predictive analytics for financial applications (e.g., risk analysis, estimation of creditworthiness, the recommendation of investment strategies, and portfolio management) [14], click-through rate (CTR) prediction [15], user recommendation systems [16], customer churn prediction [17], [18], cybersecurity [19], fraud detection [20], identity protection [21], psychology [22], delay estimations [23], anomaly detection [24], and so forth. In all these applications, a boost in predictive performance and robustness may have considerable benefits for both end users and companies that provide such solutions. Simultaneously, this requires handling many data-related pitfalls, such as noise, impreciseness, different attribute types and value ranges, or the missing value problem and privacy issues.

Meanwhile, deep neural networks offer multiple advantages over traditional machine learning methods. First, these methods are highly flexible [25], allow for efficient and iterative training, and are particularly valuable for AutoML [26]–[31]. Second, tabular data generation is possible using deep neural networks and can, for instance, help mitigate class imbalance problems [32]. Third, neural networks can be deployed for multimodal learning problems where tabular data can be one of many input modalities [28], [33]–[36], for tabular data distillation [37], [38], for federated learning [39], and in many more scenarios.

Successful deployments of data-driven applications require solving several tasks, among which we identified three *core challenges*: (1) *inference* (2) *data generation*, and (3) *interpretability*. The most crucial task is inference which is concerned with making predictions based on past observations. While a powerful predictive model is critical for all the applications mentioned in the previous paragraph, the interplay

Deep Neural Networks and Tabular Data

❖ Tabular Dataset를 활용한 Deep Learning 방법 분류



데이터 변환 방법을 사용

Tabular Data에 전문화된 구조 사용

- Hybrid: 고전적인 기계학습 + NN

- Transformer : Attention Mechanism 사용

정규화 모델

- 비선형성과 모델 복잡도 제약을 위함.

Deep Neural Networks and Tabular Data

	Method	Interpretability	Key Characteristics
Encoding	SuperTML [87]		Transform tabular data into images for CNNs
	VIME [88]		Self-supervised learning and contextual embedding
	IGTD [80]		Transform tabular data into images for CNNs
	SCARF [89]		Self-supervised contrastive learning
Architectures, Hybrid	Wide&Deep [90]		Embedding layer for categorical features
	DeepFM [15]		Factorization machine for categorical data
	SDT [91]	✓	Distill neural network into interpretable decision tree
	xDeepFM [92]		Compressed interaction network
	TabNN [93]		DNNs based on feature groups distilled from GBDT
	DeepGBM [70]		Two DNNs, distill knowlegde from decision tree
	NODE [6]		Differentiable oblivious decision trees ensemble
	NON [94]		Network-on-network model
	DNN2LR [95]		Calculate cross feature wields with DNNs for LR
	Net-DNF [57]		Structure based on disjunctive normal form
	Boost-GNN [96]		GNN on top decision trees from the GBDT algorithm
	SDTR [97]		Hierarchical differentiable neural regression model
Architectures, Transformer	TabNet [5]	✓	Sequential attention structure
	TabTransformer [98]	✓	Transformer network for categorical data
	SAINT [9]	✓	Attention over both rows and columns
	ARM-Net [99]		Adaptive relational modelling with multi-headgated attention network
	Non-Param. Transformer [100]		Process the entire dataset at once, use attention between data points
Regul.	RLN [72]	✓	Hyperparameters regularization scheme
	Regularized DNNs [10]		A "cocktail" of regularization techniques

Deep Neural Networks and Tabular Data

❖ **Deep Learning 모델별 비교를 위한 실험**

[Data Set]

	HELOC	Adult Income	HIGGS	Covertype	California Housing
Samples	9.871	32.561	11 M.	581.012	20.640
Num. features	21	6	27	52	8
Cat. features	2	8	1	2	0
Task	Binary	Binary	Binary	Multi-Class	Regression
Classes	2	2	2	7	-

Deep Neural Networks and Tabular Data

❖ 모델별 결과 비교 (Bold: Top, under line : second)

	HELOC		Adult		HIGGS		Covertype		Cal. Housing
	Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑	MSE ↓
Linear Model	73.0±0.0	80.1±0.1	82.5±0.2	85.4±0.2	64.1±0.0	68.4±0.0	72.4±0.0	92.8±0.0	0.528±0.008
KNN [65]	72.2±0.0	79.0±0.1	83.2±0.2	87.5±0.2	62.3±0.1	67.1±0.0	70.2±0.1	90.1±0.2	0.421±0.009
Decision Tree [197]	80.3±0.0	89.3±0.1	85.3±0.2	89.8±0.1	71.3±0.0	78.7±0.0	79.1±0.0	95.0±0.0	0.404±0.007
Random Forest [198]	82.1±0.2	90.0±0.2	86.1±0.2	91.7±0.2	71.9±0.0	79.7±0.0	78.1±0.1	96.1±0.0	0.272±0.006
XGBoost [53]	83.5±0.2	92.2±0.0	87.3±0.2	92.8±0.1	77.6±0.0	85.9±0.0	**97.3±0.0**	**99.9±0.0**	0.206±0.005
LightGBM [78]	83.5±0.1	92.3±0.0	**87.4±0.2**	**92.9±0.1**	77.1±0.0	85.5±0.0	93.5±0.0	99.7±0.0	**0.195±0.005**
CatBoost [79]	**83.6±0.3**	**92.4±0.1**	87.2±0.2	92.8±0.1	77.5±0.0	85.8±0.0	96.4±0.0	99.8±0.0	0.196±0.004
Model Trees [199]	82.6±0.2	91.5±0.0	85.0±0.2	90.4±0.1	69.8±0.0	76.7±0.0	-	-	0.385±0.019
MLP [200]	73.2±0.3	80.3±0.1	84.8±0.1	90.3±0.2	77.1±0.0	85.6±0.0	91.0±0.4	76.1±3.0	0.263±0.008
DeepFM [15]	73.6±0.2	80.4±0.1	86.1±0.2	91.7±0.1	76.9±0.0	83.4±0.0	-	-	0.260±0.006
DeepGBM [70]	78.0±0.4	84.1±0.1	84.6±0.3	90.8±0.1	74.5±0.0	83.0±0.0	-	-	0.856±0.065
RLN [72]	73.2±0.4	80.1±0.4	81.0±1.6	75.9±8.2	71.8±0.2	79.4±0.2	77.2±1.5	92.0±0.9	0.348±0.013
TabNet [5]	81.0±0.1	90.0±0.1	85.4±0.2	91.1±0.1	76.5±1.3	84.9±1.4	93.1±0.2	99.4±0.0	0.346±0.007
VIME [88]	72.7±0.0	79.2±0.0	84.8±0.2	90.5±0.2	76.9±0.2	85.5±0.1	90.9±0.1	82.9±0.7	0.275±0.007
TabTransformer [98]	73.3±0.1	80.1±0.2	85.2±0.2	90.6±0.2	73.8±0.0	81.9±0.0	76.5±0.3	72.9±2.3	0.451±0.014
NODE [6]	79.8±0.2	87.5±0.2	85.6±0.3	91.1±0.2	76.9±0.1	85.4±0.1	89.9±0.1	98.7±0.0	0.276±0.005
Net-DNF [57]	82.6±0.4	91.5±0.2	85.7±0.2	91.3±0.1	76.6±0.1	85.1±0.1	94.2±0.1	99.1±0.0	-
STG [201]	73.1±0.1	80.0±0.1	85.4±0.1	90.9±0.1	73.9±0.1	81.9±0.1	81.8±0.3	96.2±0.0	0.285±0.006
NAM [202]	73.3±0.1	80.7±0.3	83.4±0.1	86.6±0.1	53.9±0.6	55.0±1.2	-	-	0.725±0.022
SAINT [9]	82.1±0.3	90.7±0.2	86.1±0.3	91.6±0.2	**79.8±0.0**	**88.3±0.0**	96.3±0.1	99.8±0.0	0.226±0.004

Deep Neural Networks and Tabular Data

❖ Deep Learning 모델별 결과 비교

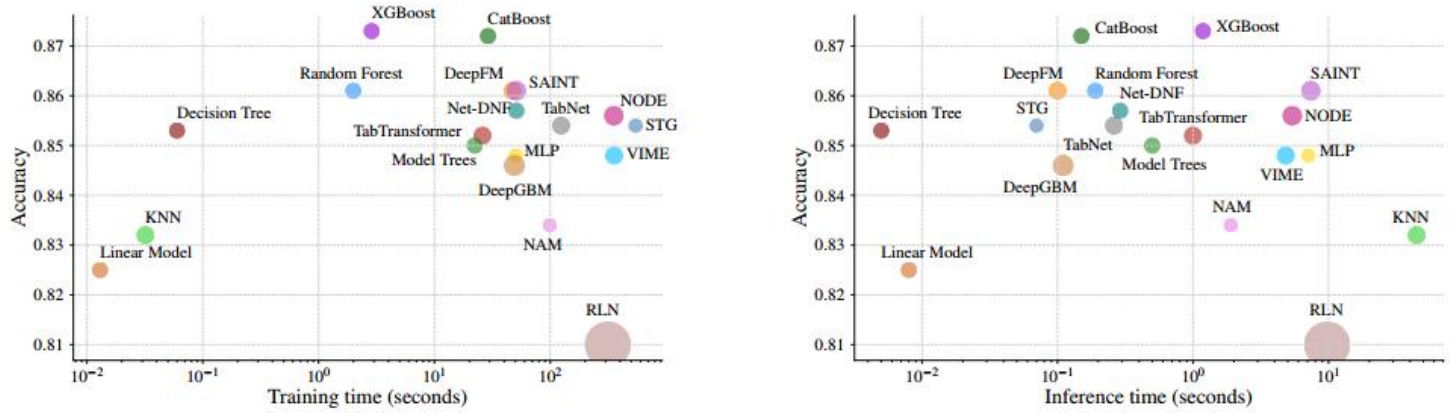


Fig. 3: Train (left) and inference (right) time benchmarks for selected methods on the Adult data set with 32.561 samples. The circle size reflects the accuracy standard deviation.

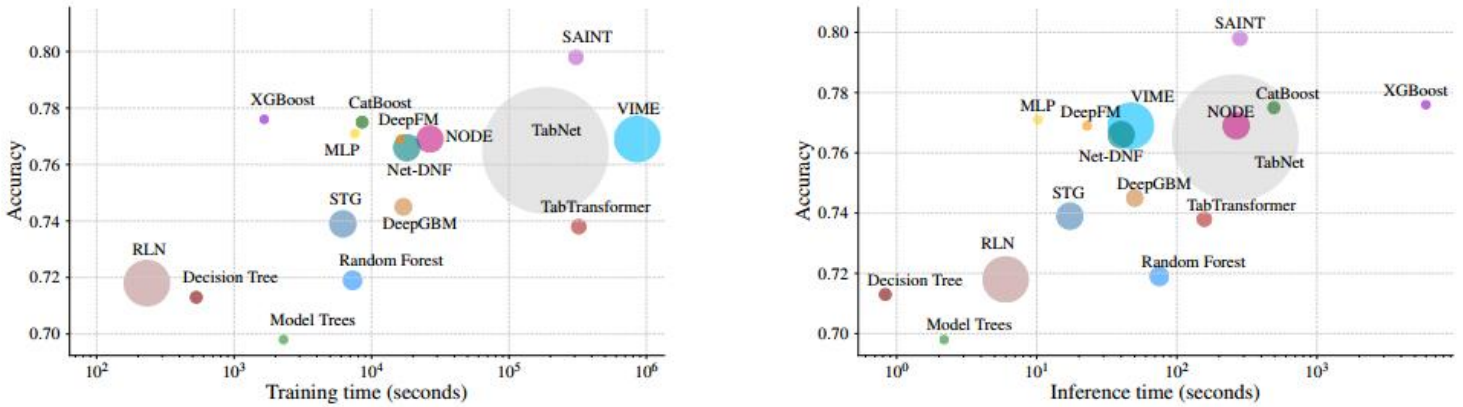


Fig. 4: Train (left) and inference (right) time benchmarks for selected methods on the HIGGS data set with 11 million samples. The circle size reflects the accuracy standard deviation.

Deep Neural Networks and Tabular Data

❖ 요약

- Tabular Data를 모델링 하는 방법을 3가지로 분류했다.

  - 데이터 변환 방법, 특화된 아키텍처 활용, 정규화 방법

- 다양한 Dataset에 Test결과 Tree 계열의 앙상블 모델은 여전히 최고의 성능

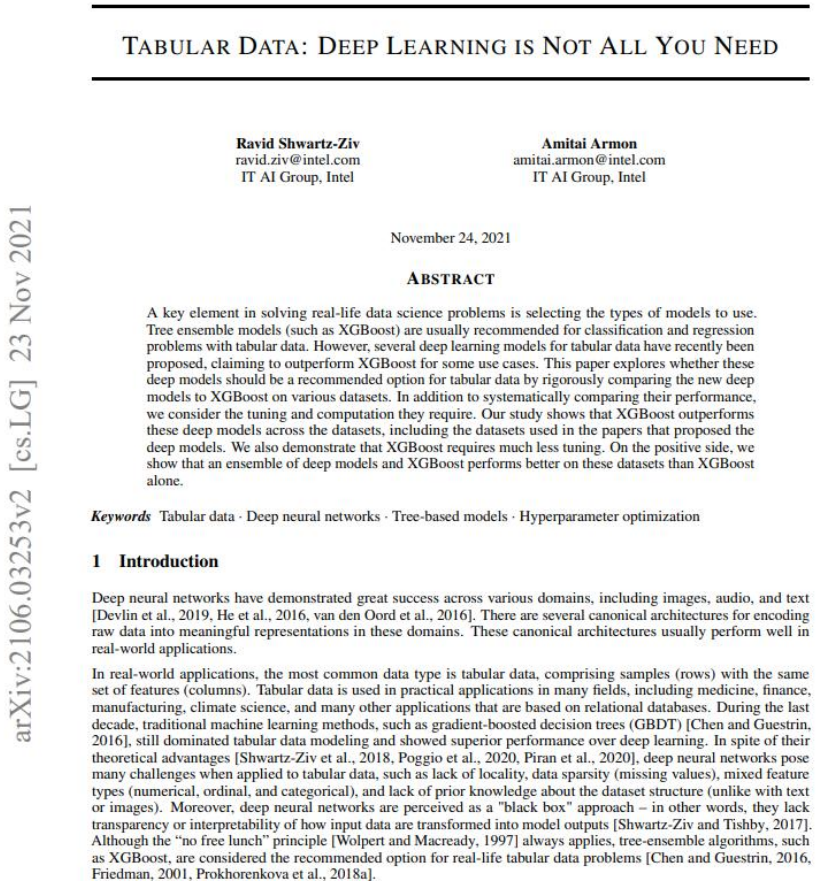
- 중소형 Dataset : Tree 기반 모델의 좋은 성능 (XGBoost, LightGBM, CatBoost)

- 대형 Dataset : SAINT가 고전적인 Machine Learning보다 좋은 성능을 보였다.

Comparison of deep learning methods for tabular dataset

❖ Paper : TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED

- Shwartz-Ziv, R., & Armon, A. (2021). *Information Fusion*, *81*, 84-90.(124회 인용)



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

ABSTRACT

A key element in solving real-life data science problems is selecting the types of models to use. Tree ensemble models (such as XGBoost) are usually recommended for classification and regression problems with tabular data. However, several deep learning models for tabular data have recently been proposed, claiming to outperform XGBoost for some use cases. This paper explores whether these deep models should be a recommended option for tabular data by rigorously comparing the new deep models to XGBoost on various datasets. In addition to systematically comparing their performance, we consider the tuning and computation they require. Our study shows that XGBoost outperforms these deep models across the datasets, including the datasets used in the papers that proposed the deep models. We also demonstrate that XGBoost requires much less tuning. On the positive side, we show that an ensemble of deep models and XGBoost performs better on these datasets than XGBoost alone.

Keywords Tabular data · Deep neural networks · Tree-based models · Hyperparameter optimization

1 Introduction

Deep neural networks have demonstrated great success across various domains, including images, audio, and text [Devlin et al., 2019, He et al., 2016, van den Oord et al., 2016]. There are several canonical architectures for encoding raw data into meaningful representations in these domains. These canonical architectures usually perform well in real-world applications.

In real-world applications, the most common data type is tabular data, comprising samples (rows) with the same set of features (columns). Tabular data is used in practical applications in many fields, including medicine, finance, manufacturing, climate science, and many other applications that are based on relational databases. During the last decade, traditional machine learning methods, such as gradient-boosted decision trees (GBDT) [Chen and Guestrin, 2016], still dominated tabular data modeling and showed superior performance over deep learning. In spite of their theoretical advantages [Shwartz-Ziv et al., 2018, Poggio et al., 2020, Piran et al., 2020], deep neural networks pose many challenges when applied to tabular data, such as lack of locality, data sparsity (missing values), mixed feature types (numerical, ordinal, and categorical), and lack of prior knowledge about the dataset structure (unlike with text or images). Moreover, deep neural networks are perceived as a "black box" approach – in other words, they lack transparency or interpretability of how input data are transformed into model outputs [Shwartz-Ziv and Tishby, 2017]. Although the "no free lunch" principle [Wolpert and Macready, 1997] always applies, tree-ensemble algorithms, such as XGBoost, are considered the recommended option for real-life tabular data problems [Chen and Guestrin, 2016, Friedman, 2001, Prokhorenkova et al., 2018a].

TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED

❖ 모델별 비교를 위한 실험

[Data Set]

Dataset	Features	Classes	Samples	Source	Paper
Gesture Phase	32	5	9.8k	OpenML	DNF-Net
Gas Concentrations	129	6	13.9k	OpenML	DNF-Net
Eye Movements	26	3	10.9k	OpenML	DNF-Net
Epsilon	2000	2	500k	PASCAL Challenge 2008	NODE
YearPrediction	90	1	515k	Million Song Dataset	NODE
Microsoft (MSLR)	136	5	964k	MSLR-WEB10K	NODE
Rossmann Store Sales	10	1	1018K	Kaggle	TabNet
Forest Cover Type	54	7	580k	Kaggle	TabNet
Higgs Boson	30	2	800k	Kaggle	TabNet
Shrutime	11	2	10k	Kaggle	New dataset
Blastchar	20	2	7k	Kaggle	New dataset

TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED

❖ 모델별 비교를 위한 실험

Model Name	Rossman	CoverType	Higgs	Gas	Eye	Gesture
XGBoost	490.18 ± 1.19	3.13 ± 0.09	21.62 ± 0.33	2.18 ± 0.20	**56.07**±0.65	80.64 ± 0.80
NODE	488.59 ± 1.24	4.15 ± 0.13	21.19 ± 0.69	2.17 ± 0.18	68.35 ± 0.66	92.12 ± 0.82
DNF-Net	503.83 ± 1.41	3.96 ± 0.11	23.68 ± 0.83	**1.44** ±0.09	68.38 ± 0.65	86.98 ± 0.74
TabNet	**485.12**±1.93	3.01 ± 0.08	**21.14**±0.20	1.92 ± 0.14	67.13 ± 0.69	96.42 ± 0.87
1D-CNN	493.81 ± 2.23	3.51 ± 0.13	22.33 ± 0.73	1.79 ± 0.19	67.9 ± 0.64	97.89 ± 0.82
Simple Ensemble	488.57 ± 2.14	3.19 ± 0.18	22.46 ± 0.38	2.36 ± 0.13	58.72 ± 0.67	89.45 ± 0.89
Deep Ensemble w/o XGBoost	489.94 ± 2.09	3.52 ± 0.10	22.41 ± 0.54	1.98 ± 0.13	69.28 ± 0.62	93.50 ± 0.75
Deep Ensemble w XGBoost	485.33 ± 1.29	**2.99** ±0.08	22.34 ± 0.81	1.69 ± 0.10	59.43 ± 0.60	**78.93** ±0.73

TabNet DNF-Net

Model Name	YearPrediction	MSLR	Epsilon	Shrutime	Blastchar
XGBoost	77.98 ± 0.11	55.43±2e-2	11.12±3e-2	13.82 ± 0.19	20.39 ± 0.21
NODE	76.39 ± 0.13	55.72±3e-2	**10.39**±1e-2	14.61 ± 0.10	21.40 ± 0.25
DNF-Net	81.21 ± 0.18	56.83±3e-2	12.23±4e-2	16.8 ± 0.09	27.91 ± 0.17
TabNet	83.19 ± 0.19	56.04±1e-2	11.92±3e-2	14.94±, 0.13	23.72 ± 0.19
1D-CNN	78.94 ± 0.14	55.97±4e-2	11.08±6e-2	15.31 ± 0.16	24.68 ± 0.22
Simple Ensemble	78.01 ± 0.17	55.46±4e-2	11.07±4e-2	13.61±, 0.14	21.18 ± 0.17
Deep Ensemble w/o XGBoost	78.99 ± 0.11	55.59±3e-2	10.95±1e-2	14.69 ± 0.11	24.25 ± 0.22
Deep Ensemble w XGBoost	**76.19** ±0.21	**55.38**±1e-2	11.18±1e-2	**13.10**±0.15	**20.18**±0.16

NODE New datasets

TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED

❖ 요약

- 본 연구에서는 딥러닝 모델들이 논문을 작성할 때, 특정 Dataset에서는 좋은 성능을 보인다는 점을 지적한다.

- 특히 XGBoost와 Deep Learning을 앙상블한 경우 가장 좋은 성능을 발휘했다.

- Tabular Dataset에 대해서 딥러닝 모델보다 XGBoost가 전반적으로 좋다.

- 향후 방향은 최적화 하기 쉽고 XGBoost와 경쟁할 수 있는 새로운 딥러닝 모델 연구 필요

Question ?

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

Comparison of deep learning methods for tabular dataset

❖ **Paper : Why do tree-based models still outperform deep learning on tabular data?**

- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). *preprint arXiv:2207.08815*. (5회인용)

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

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

Edouard Oyallon
ISIR, CNRS, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

Abstract

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 (NNs). This leads to a series of challenges which should guide researchers aiming to build tabular-specific NNs: **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.

1 Introduction

Deep learning has enabled tremendous progress for learning on image, language, or even audio datasets. On tabular data, however, the picture is muddier and ensemble models based on decision trees like XGBoost remain the go-to tool for most practitioners [Sta] and data science competitions [Kossen et al., 2021]. Indeed deep learning architectures have been crafted to create inductive biases matching invariances and spatial dependencies of the data. Finding corresponding invariances is hard in tabular data, made of heterogeneous features, small sample sizes, extreme values.

Creating tabular-specific deep learning architectures is a very active area of research (see section 2) given that tree-based models are not differentiable, and thus cannot be easily composed and jointly trained with other deep learning blocks. Most corresponding publications claim to beat or match tree-based models, but their claims have been put into question: a simple Resnet seems to be competitive with some of these new models [Gorishniy et al., 2021], and most of these methods seem to fail on new datasets [Shwartz-Ziv and Armon, 2021]. Indeed, the lack of an established benchmark for tabular data learning provides additional degrees of freedom to researchers when evaluating their method. Furthermore, most tabular datasets available online are small compared to benchmarks in other

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

❖ 모델별 비교를 위한 실험

[Data Set 45개]

A.1.1 Numerical classification

OpenML benchmark: `https://www.openml.org/search?type=benchmark&study_type=task&sort=tasks_included&id=298`

dataset_name	n_samples	n_features	original link	new link
electricity	38474	7	https://openml.org/d/151	https://www.openml.org/d/44120
covertype	566602	10	https://openml.org/d/293	https://www.openml.org/d/44121
pol	10082	26	https://openml.org/d/722	https://www.openml.org/d/44122
house_16H	13488	16	https://openml.org/d/821	https://www.openml.org/d/44123
kdd_ipums_la_97-small	5188	20	https://openml.org/d/993	https://www.openml.org/d/44124
MagicTelescope	13376	10	https://openml.org/d/1120	https://www.openml.org/d/44125
bank-marketing	10578	7	https://openml.org/d/1461	https://www.openml.org/d/44126
phoneme	3172	5	https://openml.org/d/1489	https://www.openml.org/d/44127
MiniBooNE	72998	50	https://openml.org/d/41150	https://www.openml.org/d/44128
Higgs	940160	24	https://openml.org/d/42769	https://www.openml.org/d/44129
eye_movements	7608	20	https://openml.org/d/1044	https://www.openml.org/d/44130
jannis	57580	54	https://openml.org/d/41168	https://www.openml.org/d/44131
credit	16714	10	https://www.kaggle.com/c/GiveMeSomeCredit/data?select=cs-training.csv	https://www.openml.org/d/44089
california	20634	8	https://www.dcc.fc.up.pt/ ltorgo/Regression/cal_housing.html	https://www.openml.org/d/44090
wine	2554	11	https://archive.ics.uci.edu/ml/datasets/wine+quality	https://www.openml.org/d/44091

A.1.2 Numerical regression

OpenML benchmark: `https://www.openml.org/search?type=benchmark&study_type=task&sort=tasks_included&id=297`

dataset_name	n_samples	n_features	original link	new link
cpu_act	8192	21	https://openml.org/d/197	https://www.openml.org/d/44132
pol	15000	26	https://openml.org/d/201	https://www.openml.org/d/44133
elevators	16599	16	https://openml.org/d/216	https://www.openml.org/d/44134
isolet	7797	613	https://openml.org/d/300	https://www.openml.org/d/44135
wine_quality	6497	11	https://openml.org/d/287	https://www.openml.org/d/44136
Ailerons	13750	33	https://openml.org/d/296	https://www.openml.org/d/44137
houses	20640	8	https://openml.org/d/537	https://www.openml.org/d/44138
house_16H	22784	16	https://openml.org/d/574	https://www.openml.org/d/44139
diamonds	53940	6	https://openml.org/d/42225	https://www.openml.org/d/44140
Brazilian_houses	10692	8	https://openml.org/d/42688	https://www.openml.org/d/44141
Bike_Sharing_Demand	17379	6	https://openml.org/d/42712	https://www.openml.org/d/44142
nyc-taxi-green-dec-2016	581835	9	https://openml.org/d/42729	https://www.openml.org/d/44143
house_sales	21613	15	https://openml.org/d/42731	https://www.openml.org/d/44144
sulfur	10081	6	https://openml.org/d/23515	https://www.openml.org/d/44145
medical_charges	163065	5	https://openml.org/d/42720	https://www.openml.org/d/44146
MiamiHousing2016	13932	14	https://openml.org/d/43093	https://www.openml.org/d/44147
superconduct	21263	79	https://openml.org/d/43174	https://www.openml.org/d/44148
california	20640	8	https://www.dcc.fc.up.pt/ ltorgo/Regression/cal_housing.html	https://www.openml.org/d/44025
fifa	18063	5	https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset	https://www.openml.org/d/44026
year	515345	90	https://archive.ics.uci.edu/ml/datasets/yearpredictionmsd	https://www.openml.org/d/44027

A.1.3 Categorical classification

OpenML benchmark: `https://www.openml.org/search?type=benchmark&sort=date&study_type=task&id=300`

dataset_name	n_samples	n_features	Original link	New link
electricity	38474	8	https://openml.org/d/151	https://www.openml.org/d/44156
eye_movements	7608	23	https://openml.org/d/1044	https://www.openml.org/d/44157
KDDCup09_upselling	5032	45	https://openml.org/d/1114	https://www.openml.org/d/44158
covertype	423680	54	https://openml.org/d/1596	https://www.openml.org/d/44159
rl	4970	12	https://openml.org/d/41160	https://www.openml.org/d/44160
road-safety	111762	32	https://openml.org/d/42803	https://www.openml.org/d/44161
compass	16644	17	https://www.kaggle.com/datasets/danofer/compass?select=cox-violent-parsed.csv	https://www.openml.org/d/44162

A.1.4 Categorical regression

OpenML benchmark: `https://www.openml.org/search?type=benchmark&study_type=task&sort=tasks_included&id=299`

dataset_name	n_features	n_samples	Original link	New link
yprop_4_1	62	8885	https://openml.org/d/416	https://www.openml.org/d/44054
analcatdata_supreme	7	4052	https://openml.org/d/504	https://www.openml.org/d/44055
visualizing_soil	4	8641	https://openml.org/d/688	https://www.openml.org/d/44056
black_friday	9	166821	https://openml.org/d/41540	https://www.openml.org/d/44057
diamonds	9	53940	https://openml.org/d/42225	https://www.openml.org/d/44059
Mercedes_Benz_Greener_Manufacturing	359	4209	https://openml.org/d/42570	https://www.openml.org/d/44061
Brazilian_houses	11	10692	https://openml.org/d/42688	https://www.openml.org/d/44062
Bike_Sharing_Demand	11	17379	https://openml.org/d/42712	https://www.openml.org/d/44063
OnlineNewsPopularity	59	39644	https://openml.org/d/42724	https://www.openml.org/d/44064
nyc-taxi-green-dec-2016	16	581835	https://openml.org/d/42729	https://www.openml.org/d/44065
house_sales	17	21613	https://openml.org/d/42731	https://www.openml.org/d/44066
particulate-matter-ukair-2017	6	394299	https://openml.org/d/42207	https://www.openml.org/d/44068
SGEMM_GPU_kernel_performance	9	241600	https://openml.org/d/43144	https://www.openml.org/d/44069

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

❖ 모델별 비교를 위한 실험

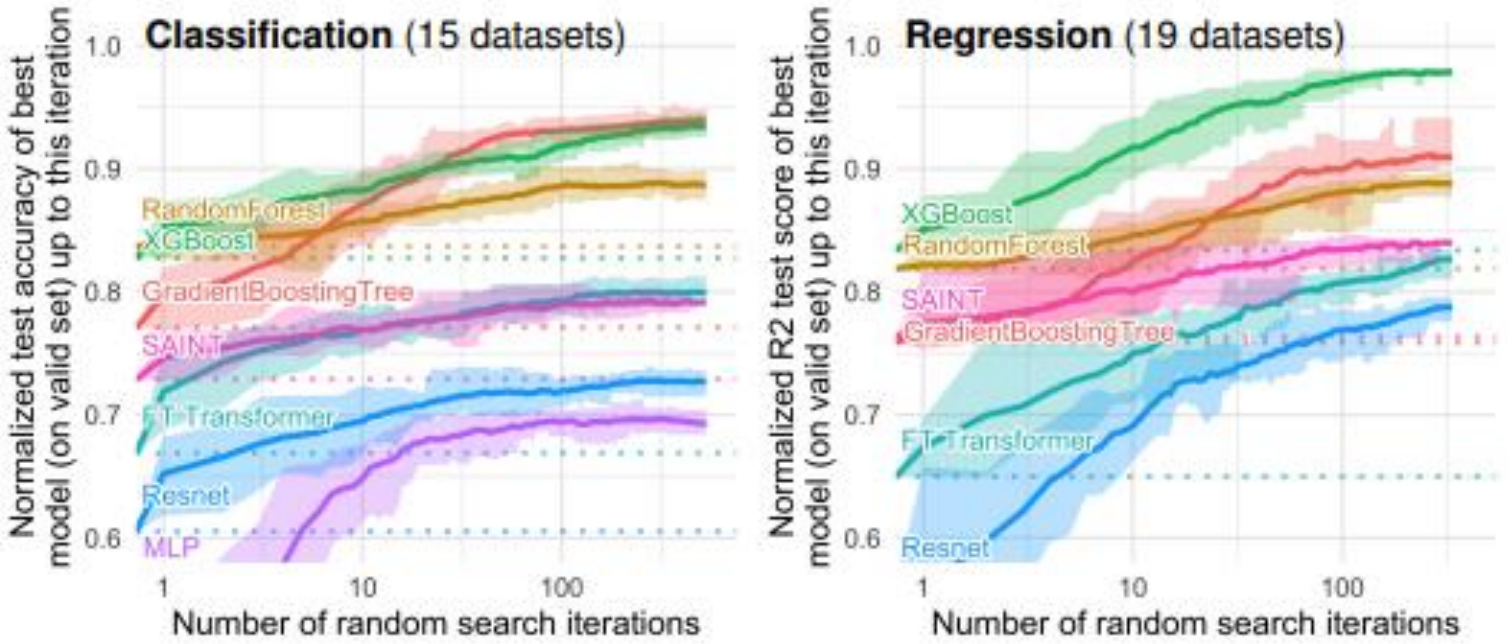


Figure 1: **Benchmark on medium-sized datasets, with only numerical features**. Dotted lines correspond to the score of the default hyperparameters, which is also the first random search iteration. Each value corresponds to the test score of the best model (on the validation set) after a specific number of random search iterations, averaged on 15 shuffles of the random search order. The ribbon corresponds to the minimum and maximum scores on these 15 shuffles.

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

❖ 모델별 비교를 위한 실험

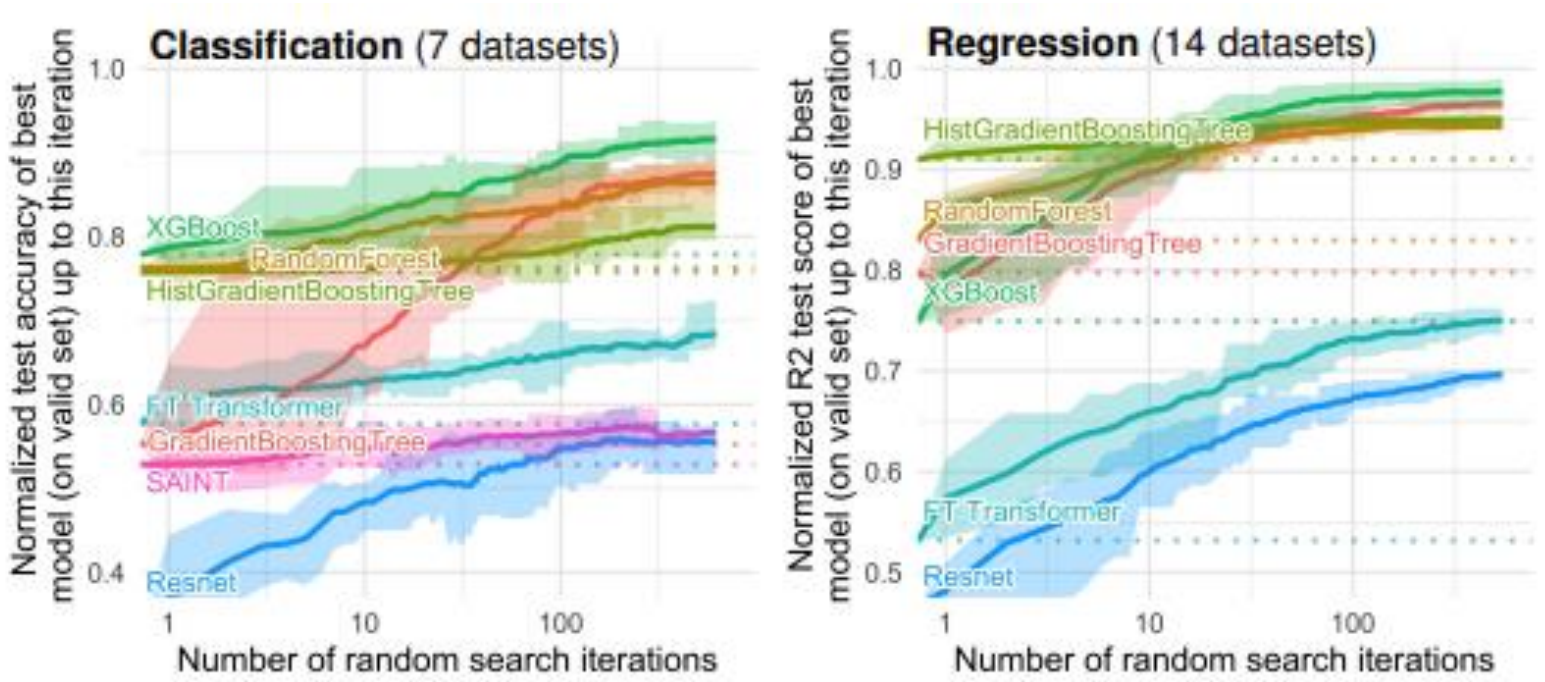
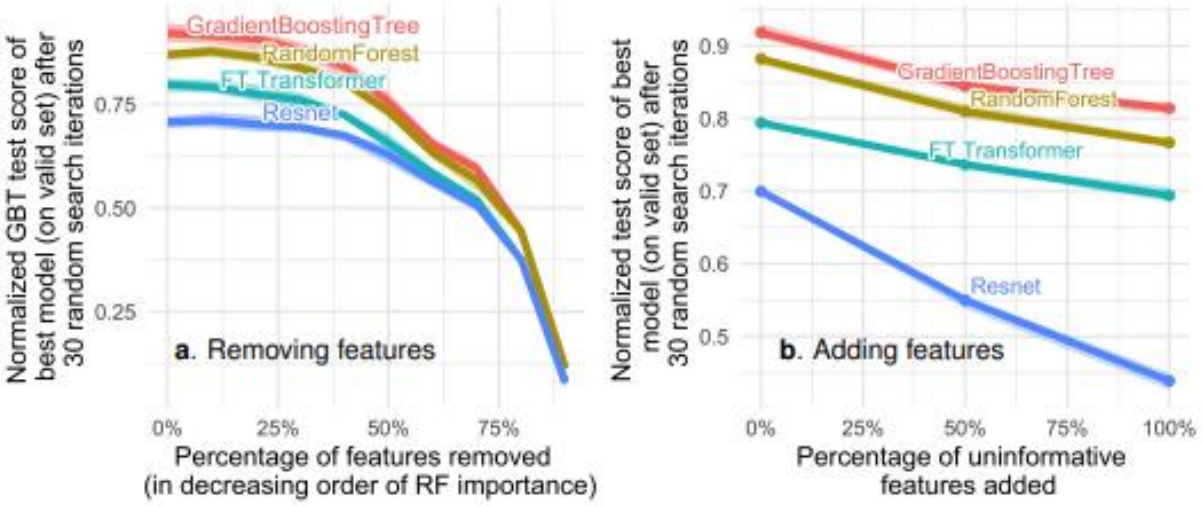


Figure 2: **Benchmark on medium-sized datasets, with both numerical and categorical features.** Dotted lines correspond to the score of the default hyperparameters, which is also the first random search iteration. Each value corresponds to the test score of the best model (on the validation set) after a specific number of random search iterations, averaged on 15 shuffles of the random search order. The ribbon corresponds to the minimum and maximum scores on these 15 shuffles.

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

❖ 요약

- 본 연구에서는 왜 Tabular Dataset에 대해서 Tree 모델이 좋은 성능을 보이는지를 고찰했다.
- (1) Neural Network 모델의 경우 편향이 심하다.
- (2) Tabular Data에는 의미 없는 Feature가 있고 Neural Network에 영향을 미친다. Tree 계열 모델들은 의미 없는 Feature들에 대해 Robust하다.



Conclusion

❖ 결론

- 이번 Seminar는 Tabular Dataset에 대한 딥러닝 방법론들과 고전적인 기계학습 방법들을 비교하는 논문 3편을 리뷰

- 3편의 논문들은 유사한 방식으로 모델을 비교
  - 동일한 Dataset에 대해서 기존 기계학습 모델과 Deep Learning 방법들 비교
  - 일부 데이터를 제외하고 전반적으로 좋은 성능을 나타내는 것은 tree 계열 모델들

- 왜 tree계열 모델들이 딥러닝보다 성능이 좋은지에 대한 고찰
  - Neural Network의 경우 편향이 심하다.
  - Tree 계열의 모델이 정보가 없는 Feature들에 대해 더 Robust 하다.

참고자료

Deep Neural Networks and Tabular Data: A Survey
Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug,
Martin Pawelczyk and Gjergji Kasneci

Tabular Data: Deep Learning is Not All You Need
Ravid Shwartz-Ziv, Amitai Armon

Why do tree-based models still outperform deep learning on tabular data?
Léo Grinsztajn (SODA), Edouard Oyallon (ISIR, CNRS), Gaël Varoquaux (SODA)