TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED

Ravid Shwartz-Ziv - Amitai Armon 2021

Groupe De Lecture - 25/05/2022

Tabular data

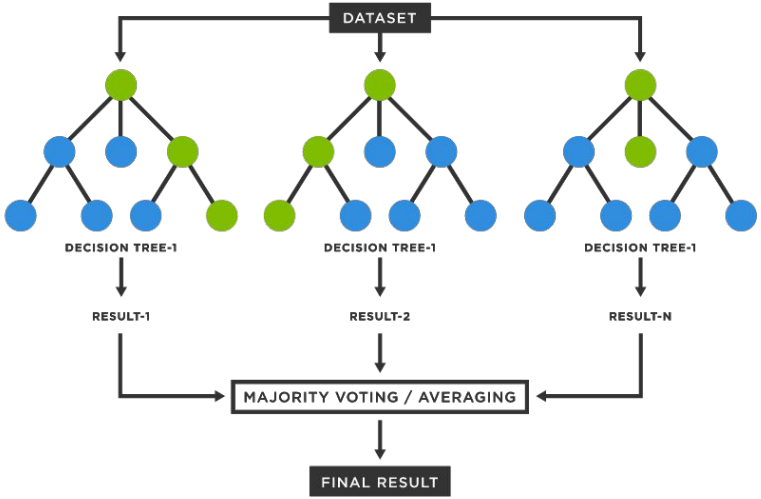
	age (n)	job (c)	marital (c)	education (c)	balance (n)	housing (c)
0	30	unemployed	married	primary	1787	no
1	33	services	married	secondary	4789	yes
2	35	management	single	tertiary	1350	yes
3	30	management	married	tertiary	1476	yes
4	59	blue-collar	married	secondary	0	yes
5	35	management	single	tertiary	747	no

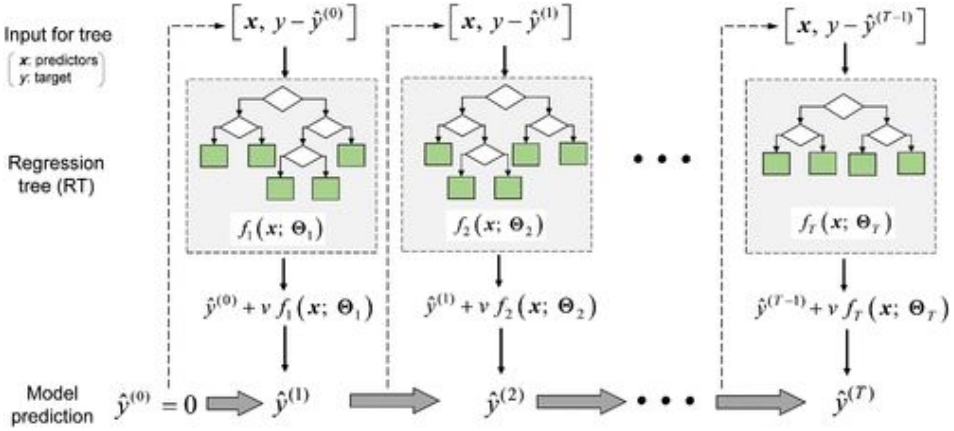
Tabular data are organized by column. Data are mainly categorical (binary or multiple classes) or numerical.

Tree based models

Random forest :

On construit des arbres sur des parties différentes des données et on agrège.

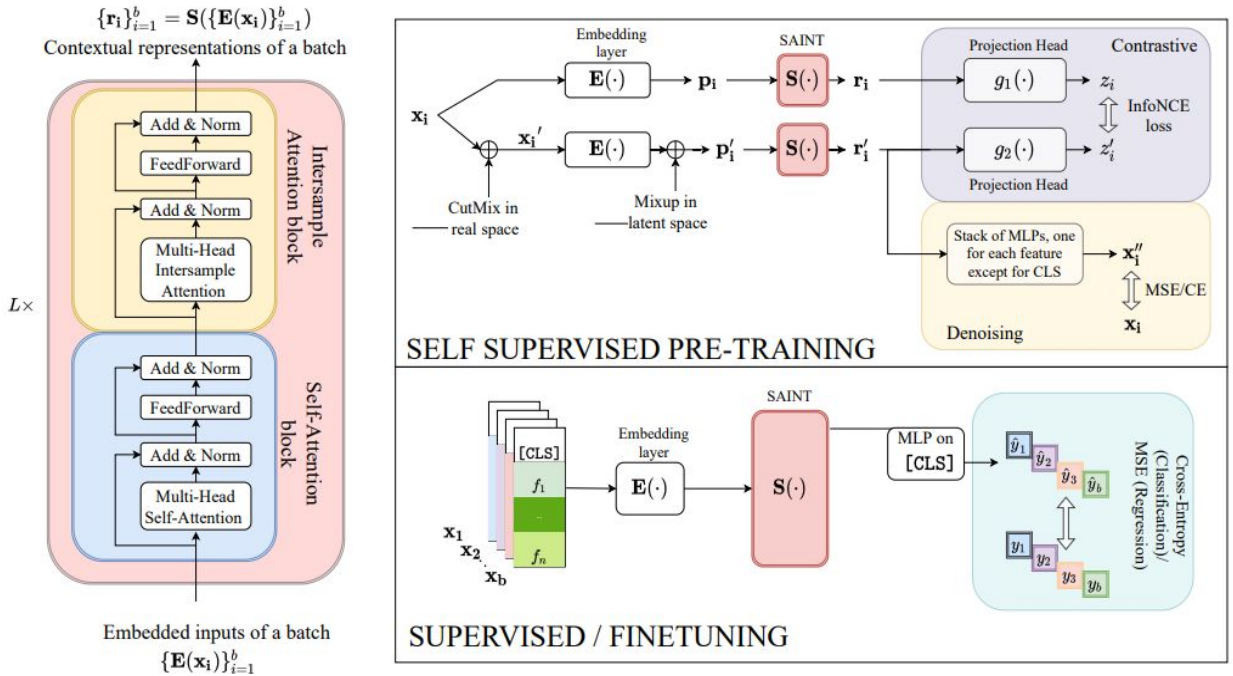




XGBoost :

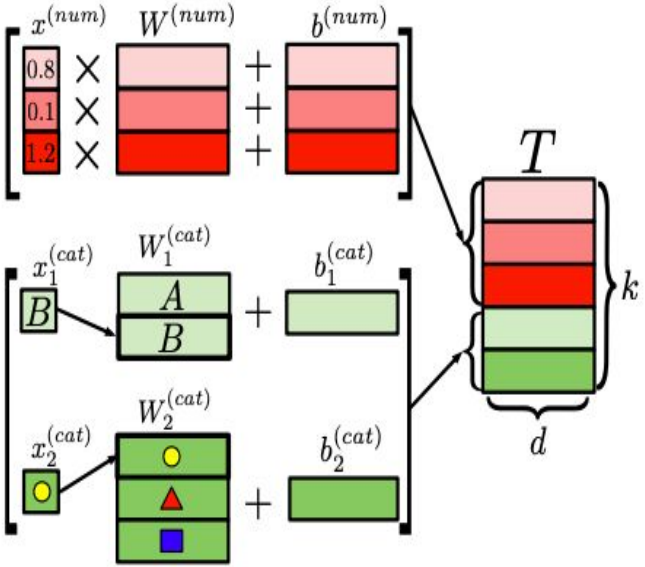
On construit des arbres successivement à prédire les erreurs des précédents.
Techniquement, il s'agit de faire une méthode de Newton dont les pas sont calculés par les arbres de décisions.

Best deep learning model for tabular data

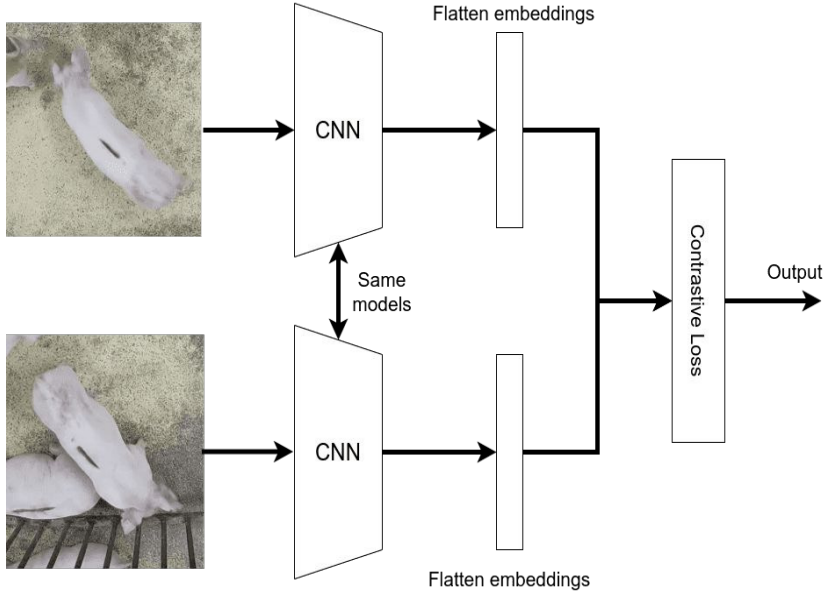


Self-Attention and Intersample Attention Transformer (SAINT)

Details

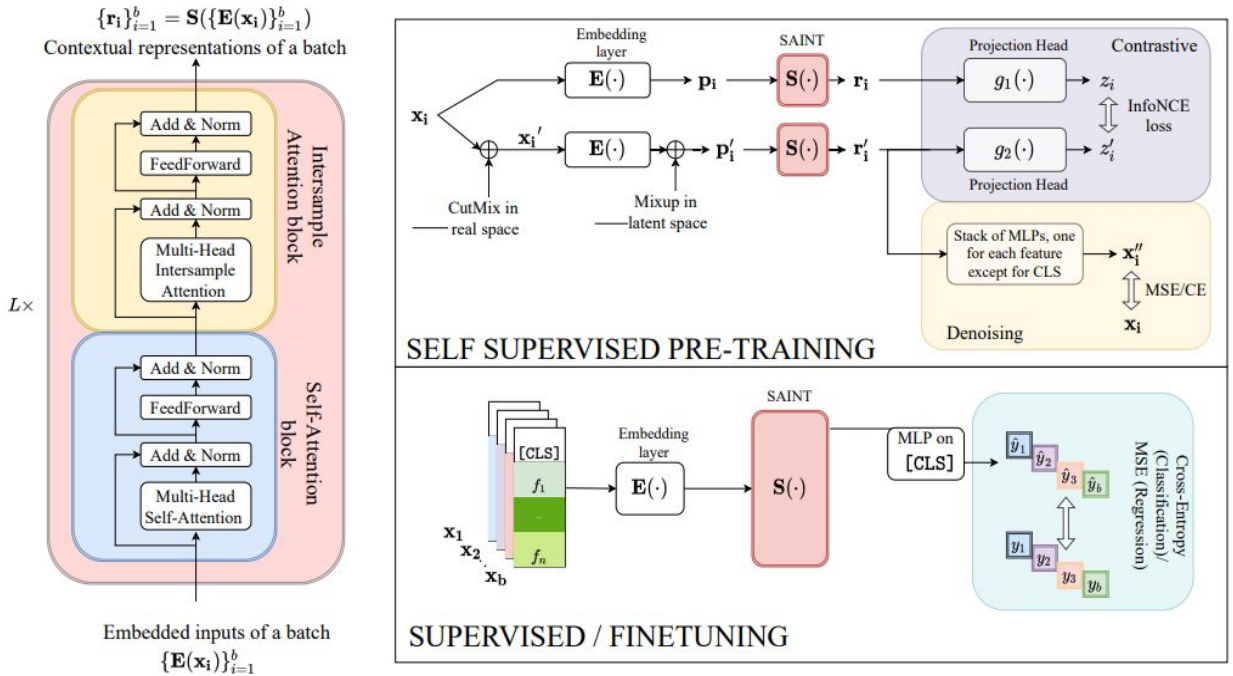


Embedding



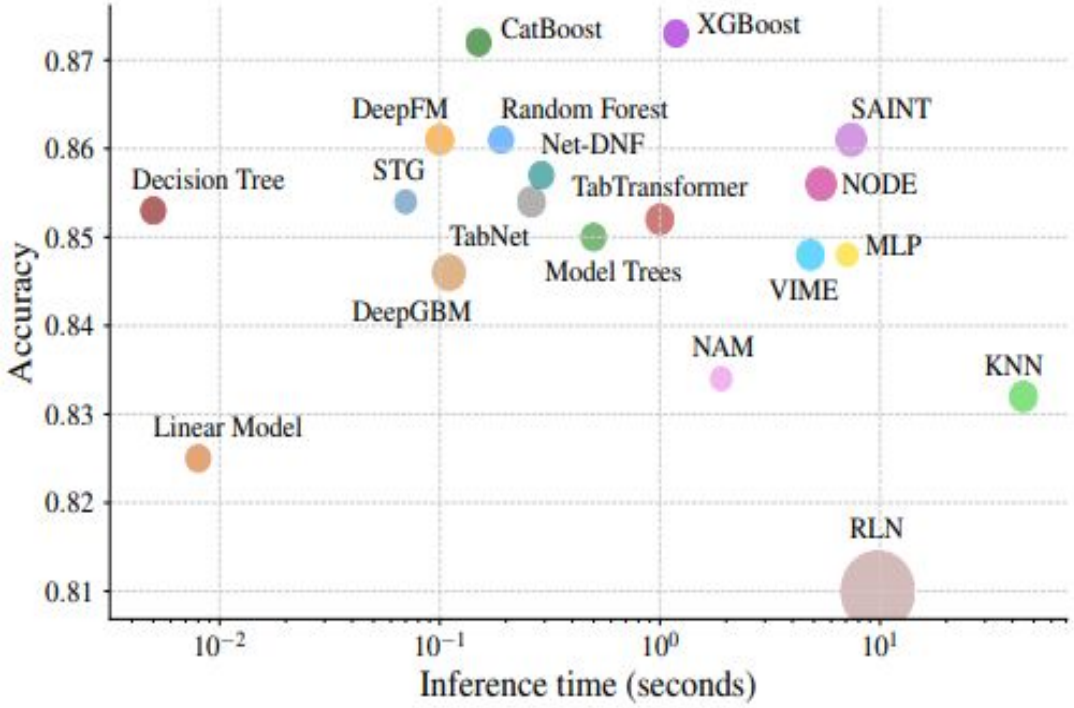
Siamese networks

Best deep learning model for tabular data



Self-Attention and Intersample Attention Transformer (SAINT)

Le constat :



[Borisov et al. 2022]

Le constat :

- In most cases, the models perform worse on unseen datasets than do the datasets' original models.
- The XGBoost model generally outperformed the deep models. For 8 of the 11 datasets, XGBoost outperformed the deep models, which did not appear in the original paper. For these datasets, the results were significant ($p < 0.005$).
- No deep model consistently outperformed the others. Each deep model was better only on the datasets that appeared in its own paper. However, the performance of the 1D-CNN model may seem better since all datasets were new to it.
- The ensemble of deep models and XGBoost outperformed the other models in most cases. For 7 of the 11 datasets, the ensemble of deep models or XGBoost was significantly better than the single deep models. The p-value in these cases was less than 0.005, which indicates the null hypothesis (i.e., no difference between the performance of the tested models) is rejected.

Pourquoi ?

Les arbres comme les réseaux sont des approximateurs universels, ils ont la même expressivité :

Pourquoi les arbres semblent-ils être meilleurs ?

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

Léo Grinsztajn - Edouard Oyallon - Gaël Varoquaux

Groupe De Lecture - 25/05/2022

Benchmarks

45 Datasets :

- Colonnes hétérogènes
- Pas trop hautement dimensionnel (d/N < 1/10)
- I.I.D
- Real-World (pas de dataset synthétiques)
- Pas trop petits (d >= 4, N >= 3000)
- Pas trop faciles (régression linéaire performe à 5% de XGBoost)
- Pas déterministe (avec du bruit)

Traitement :

- Tronqués à N = 10000
- Suppression des données manquantes
- Feature catégoriques < 20 items, Feature numériques > 10 valeurs
- Balance des classes

Benchmarks

Les difficultés :

- Tronqués à N = 10000
- Suppression des données manquantes
- Feature catégoriques < 20 items, Feature numériques > 10 valeurs
- Balance des classes

Traitement des données :

- OneHotEncoding pour les modèles ne supportant pas les variable catégoriques
- Gaussianisation des features pour les réseaux

Even more benchmarks

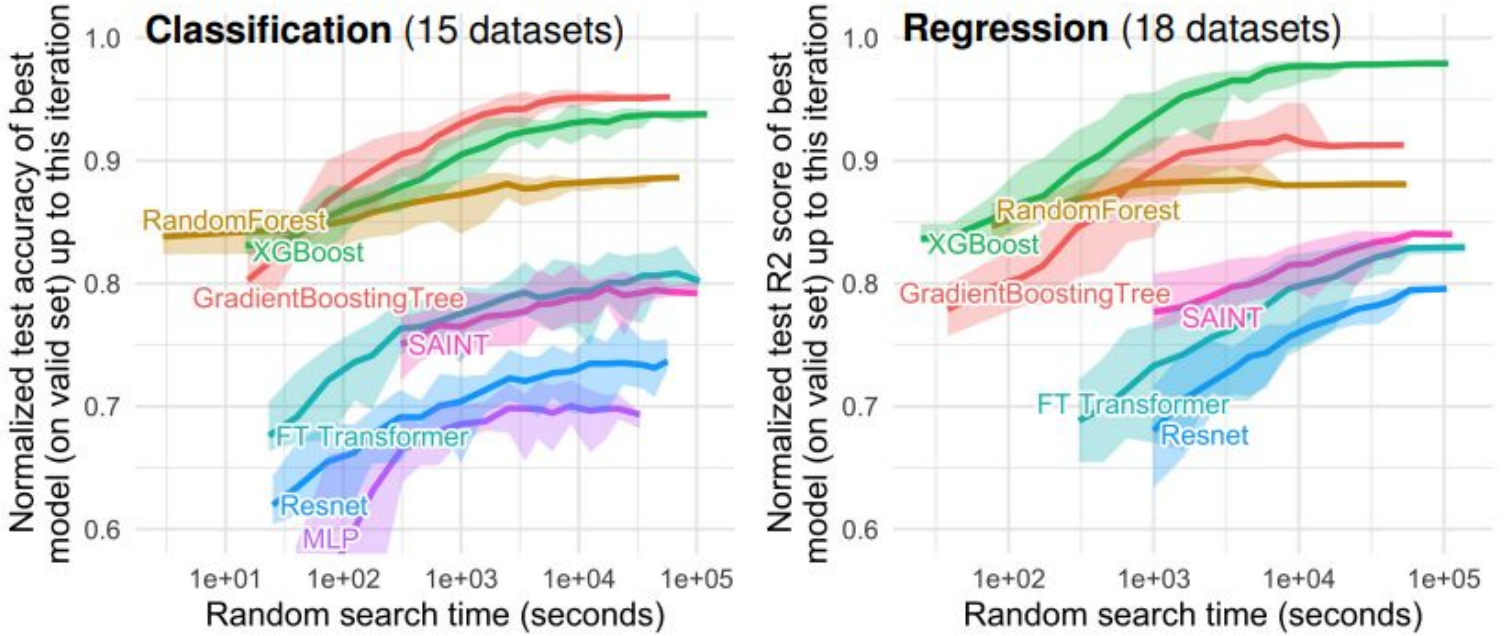


Figure 7: **Time benchmark on medium-sized datasets, with only numerical features**. The first

Even more benchmarks

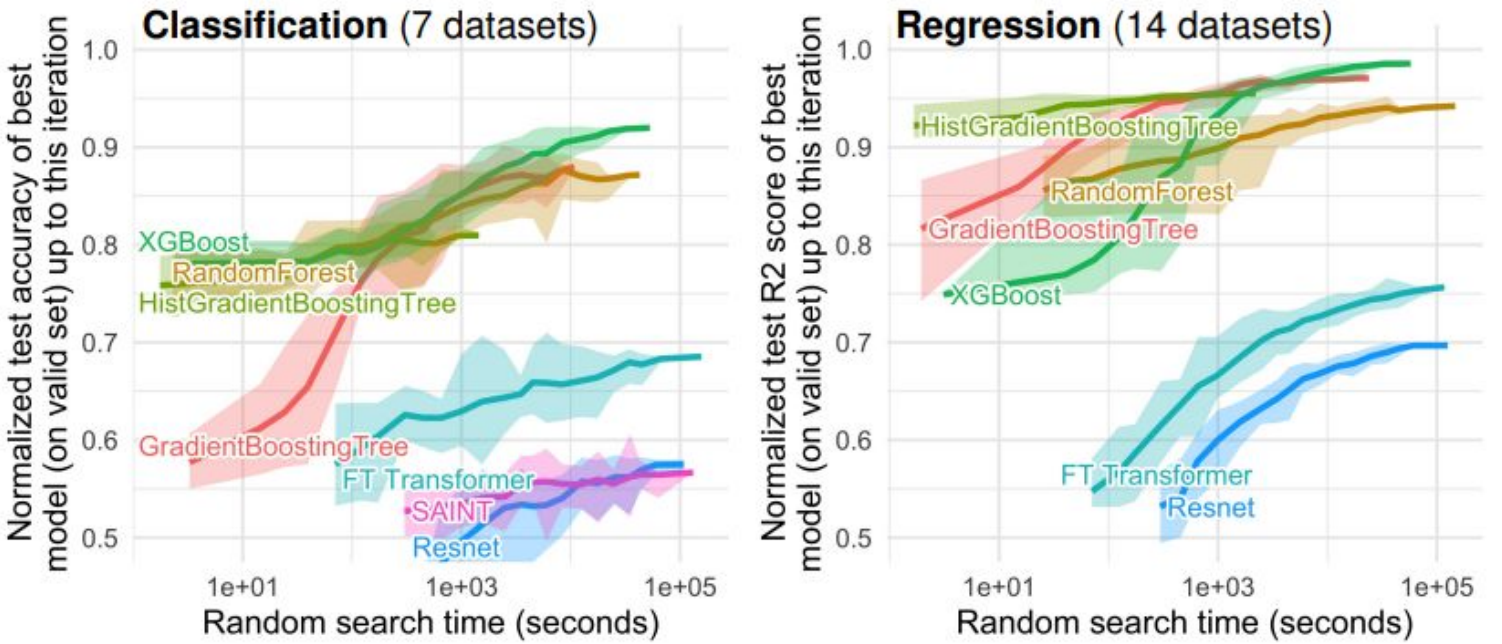
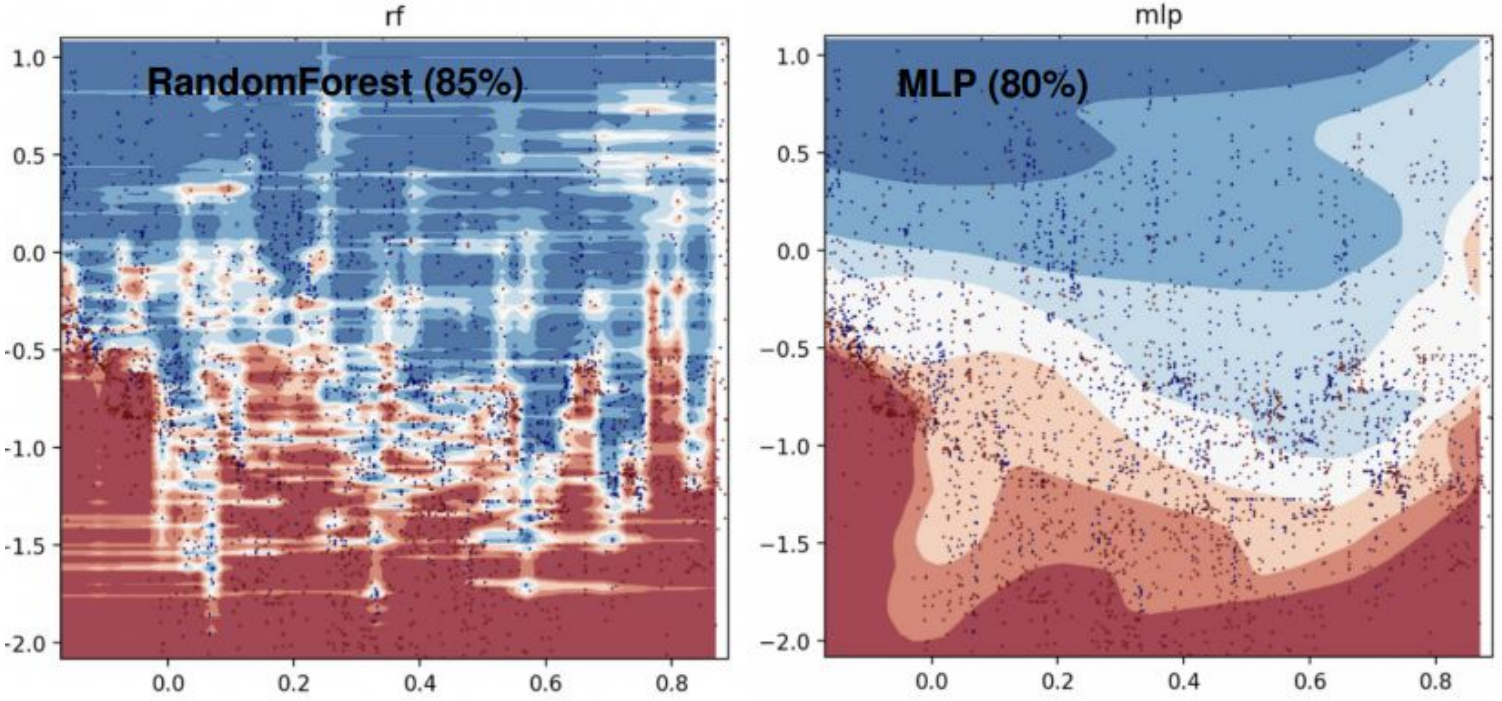


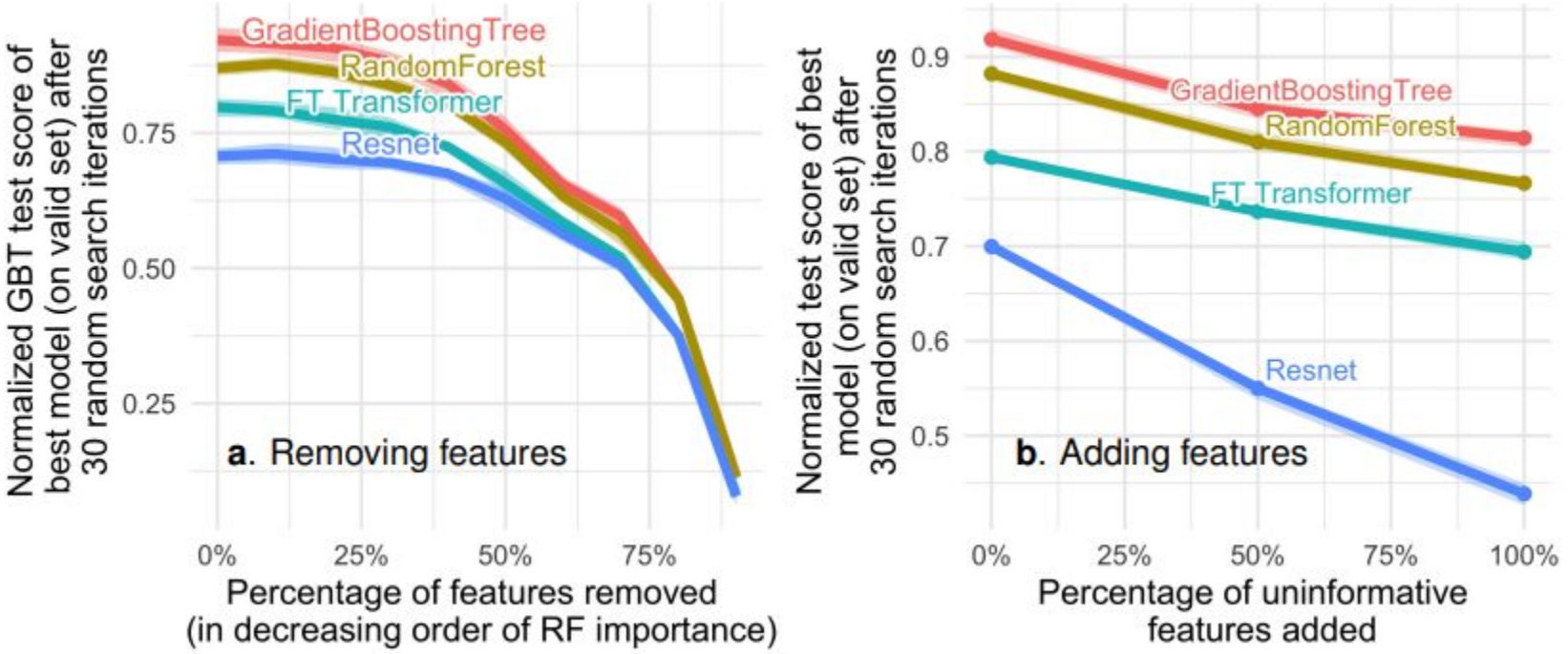
Figure 8: **Time benchmark on medium-sized datasets, with both numerical and categorical**

Finding 1: NN are biases towards smoothness

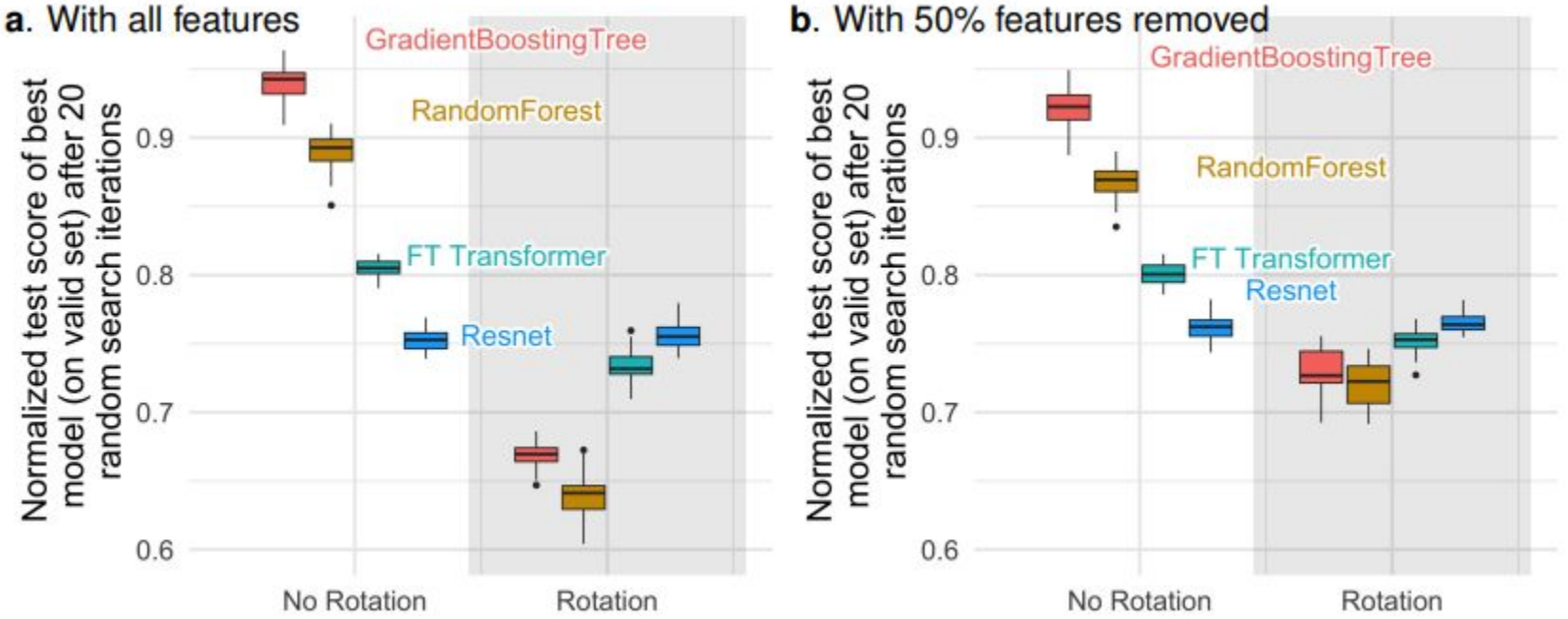


Decision boundaries

Finding 2: NN hate uninformative features



Finding 3: NN are “rotationally” invariant, but not the data…



Références

- Ravid Shwartz-Ziv, & Amitai Armon (2021). Tabular Data: Deep Learning is Not All You Need. *CoRR, abs/2106.03253.*
- Grinsztajn, L., Oyallon, E., & Varoquaux, G.. (2022). Why do tree-based models still outperform deep learning on tabular data?.
- Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, & Tom Goldstein (2021). SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. *CoRR, abs/2106.01342.*
- Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, & Gjergji Kasneci (2021). Deep Neural Networks and Tabular Data: A Survey. *CoRR, abs/2110.01889.*
- Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. doi: 10.1023/A:1010933404324
- Tianqi Chen, & Carlos Guestrin (2016). XGBoost: A Scalable Tree Boosting System. *CoRR, abs/1603.02754.*
- Ng, A. (2004). Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning* (pp. 78). Association for Computing Machinery.
- Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, & Artem Babenko (2021). Revisiting Deep Learning Models for Tabular Data. *CoRR, abs/2106.11959.*