Machine Learning for econometrics

Flexible models for tabular data

Matthieu Doutreligne

February 18th, 2025

Reminder from previous session

- Statistical learning 101: bias-variance trade-off
- Regularization for linear models: Lasso, Ridge, Elastic Net
- Transformation of variables: polynomial regression
-

Reminder from previous session

- Statistical learning 101: bias-variance trade-off
- Regularization for linear models: Lasso, Ridge, Elastic Net
- Transformation of variables: polynomial regression
- 🤔 But… How to select the best model? the best hyperparameters?

Table of contents

Model evaluation and selection with cross-validation

Example with the Wage dataset

- Raw dataset: (N=534, p=11)

EDUCATION	SOUTH	SEX	EXPERIENCE	UNION	WAGE	AGE	RACE	OCCUPATION	SECTOR	MARR
8	no	female	21	not_member	5.10	35	Hispanic	Other	Manufacturing	Married
9	no	female	42	not_member	4.95	57	White	Other	Manufacturing	Married
12	no	male	1	not_member	6.67	19	White	Other	Manufacturing	Unmarried
12	no	male	4	not_member	4.00	22	White	Other	Other	Unmarried
12	no	male	17	not_member	7.50	35	White	Other	Other	Married

-

-

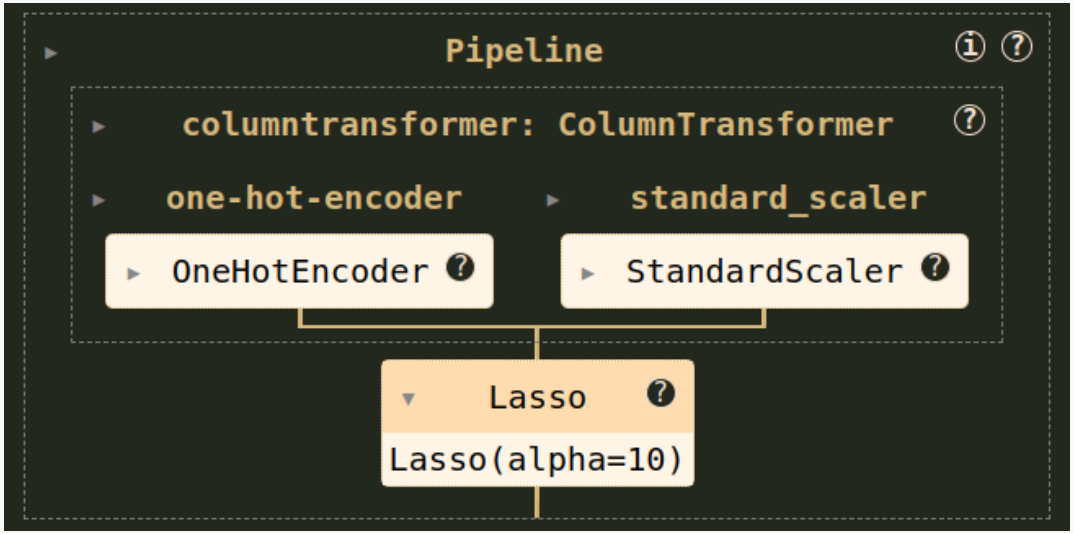
Example with the Wage dataset

- Raw dataset: (N=534, p=11)
- Transformation: encoding of categorical data and scaling of numerical data

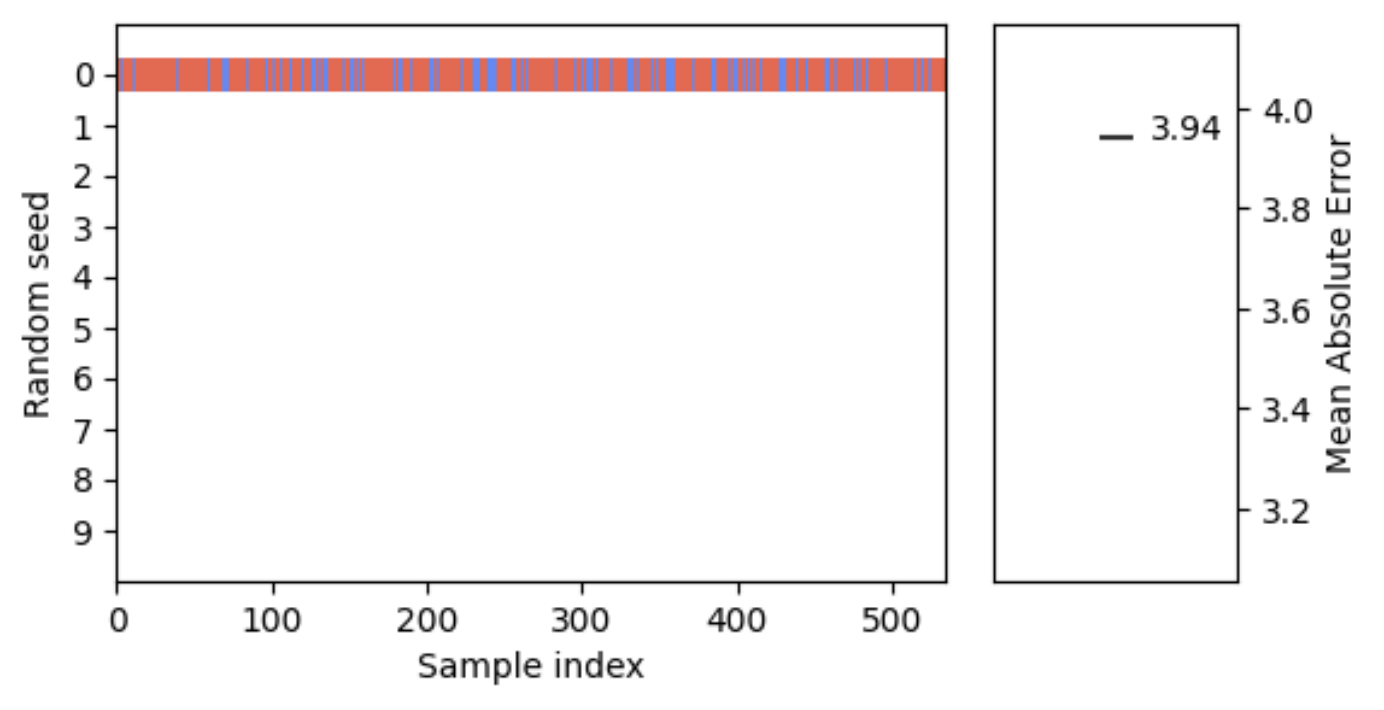
one-hot-encoder__SOUTH_no	one-hot-encoder__SOUTH_yes	one-hot-encoder__SEX_female	one-hot-encoder__SEX_male	one-hot-encoder__UNION_member	encoder__UNION_not
1.0	0.0	1.0	0.0	0.0	
1.0	0.0	1.0	0.0	0.0	
1.0	0.0	0.0	1.0	0.0	
1.0	0.0	0.0	1.0	0.0	
1.0	0.0	0.0	1.0	0.0	

Example with the Wage dataset

- Raw dataset: (N=534, p=11)
- Transformation: encoding of categorical data and scaling of numerical data
- Regressor: Lasso with regularization parameter ($\alpha = 10$), the final pipeline is:

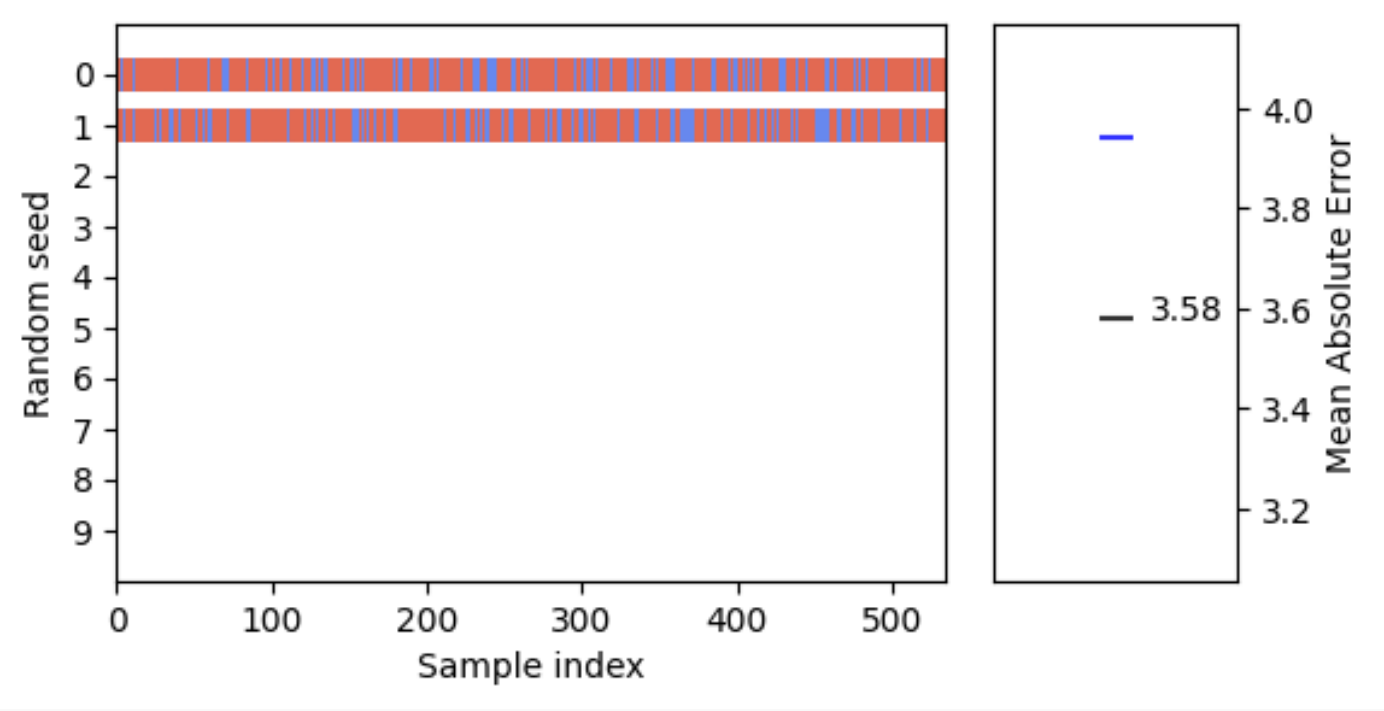


Splitting once: In red, the training set, in blue, the test set

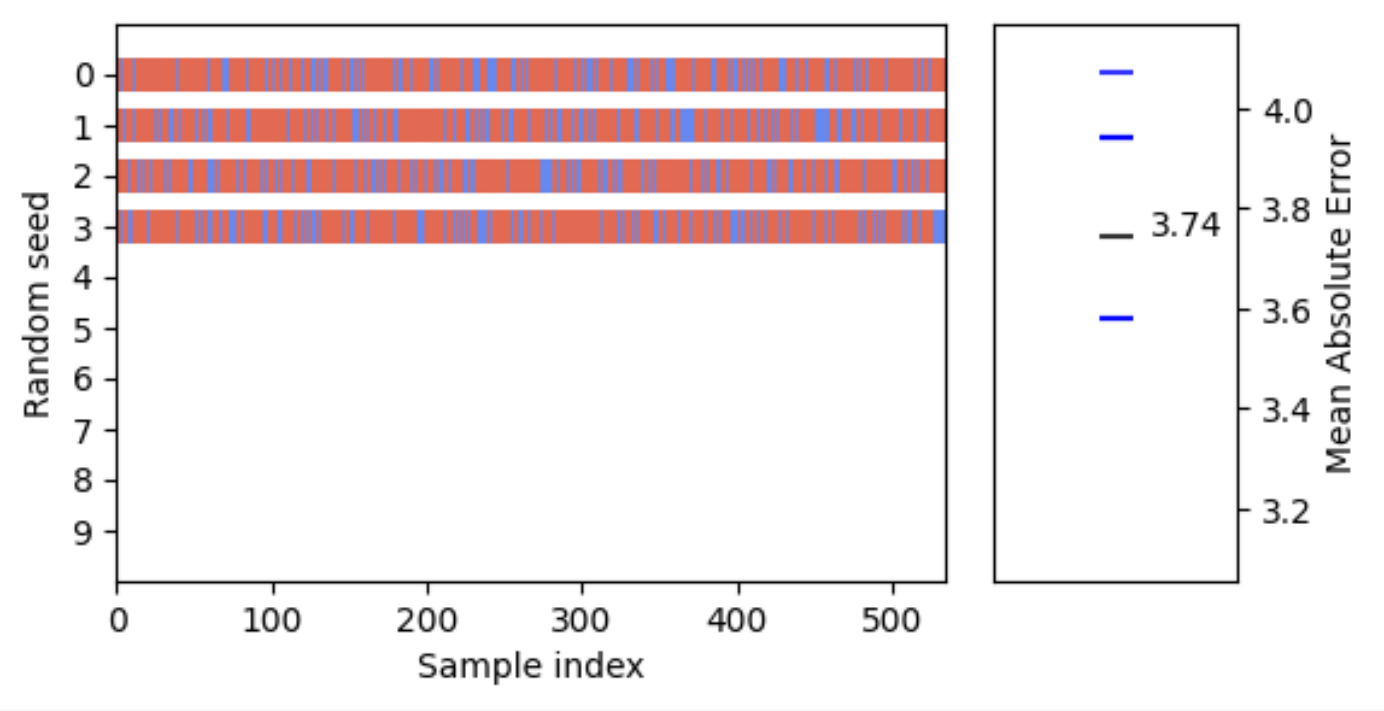


Repeated train/test splits

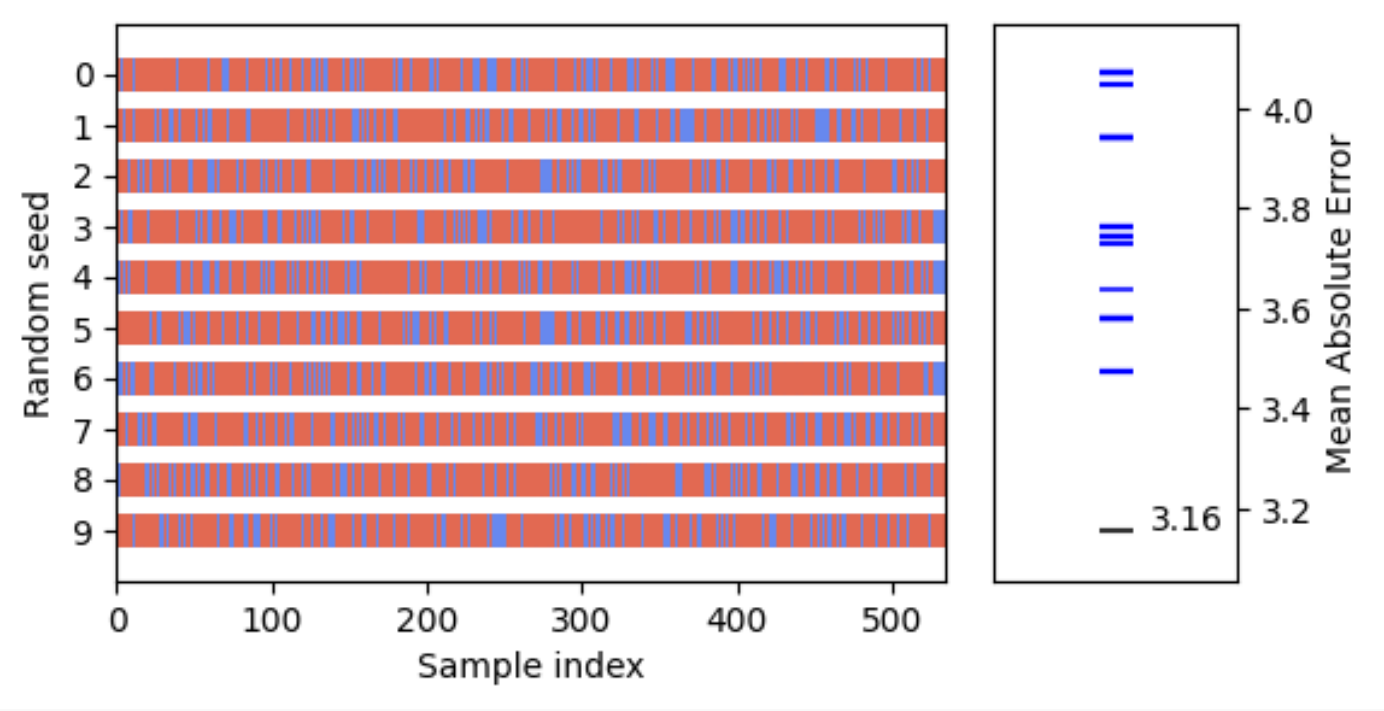
But we could have chosen another split ! Yielding a different MAE



And another split...



Splitting ten times



Repeated train/test splits = Cross-validation

- In sklearn, it can be instantiated with `cross_validate`.

```python
from sklearn.model_selection import cross_validate
from sklearn.model_selection import ShuffleSplit

cv = ShuffleSplit(n_splits=40, test_size=0.3, random_state=0)
cv_results = cross_validate(
    regressor, data, target, cv=cv, scoring="neg_mean_absolute_error"
)
```

Repeated train/test splits = Cross-validation

- In sklearn, it can be instantiated with `cross_validate`.
- It is a more robust way to evaluate the model's performance:
- We get a more robust estimate by taking the mean over the repetitions
- We get a better idea of the variability of the model's performance: similar to bootstrapping (but different)

How to select a model?

Tree, random forests and boosting

Random Forests for predictive inference

Boosting

Ensemble models

A word on other families of models

Why not use deep learning everywhere?

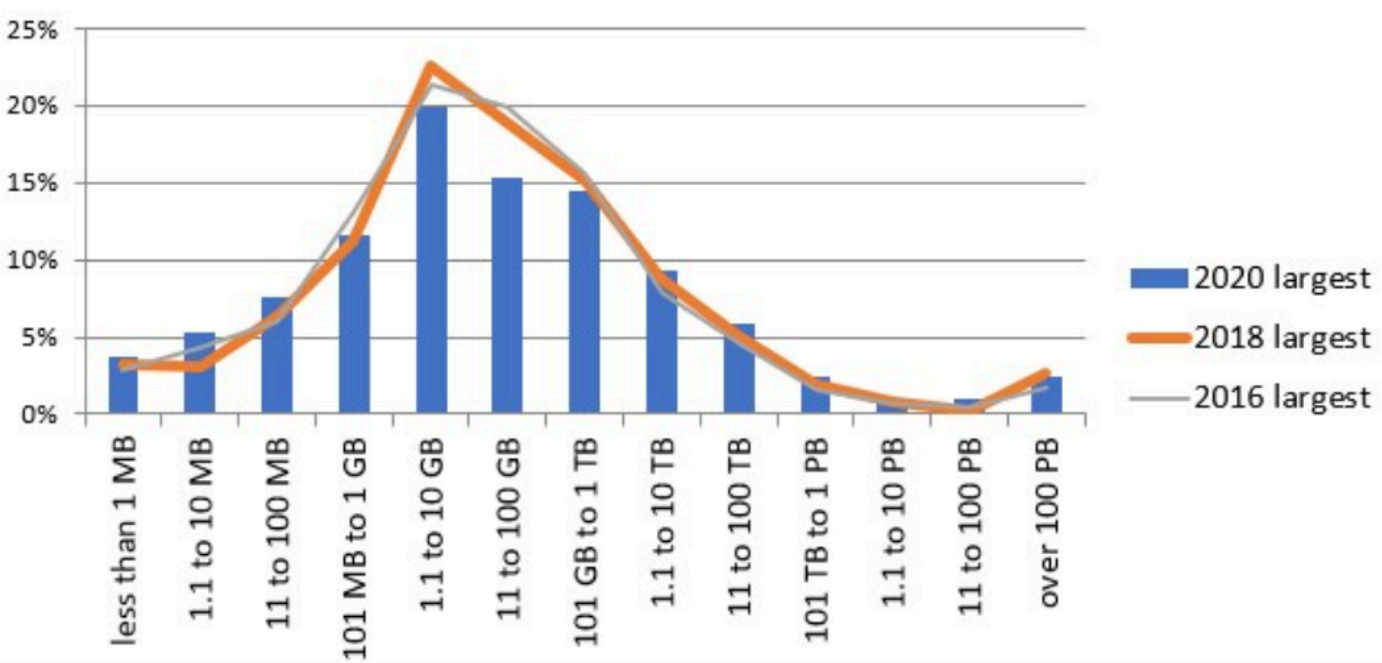
- Success of deep learning (aka deep neural networks) in image, speech recognition and text
- Why not so used in econometrics?

Deep learning needs a lot of data (typically $N \approx 1$ million)

- Do we have this much data in econometrics?

Limited data settings

- Typically in economics (but also everywhere), we have a limited number of observations

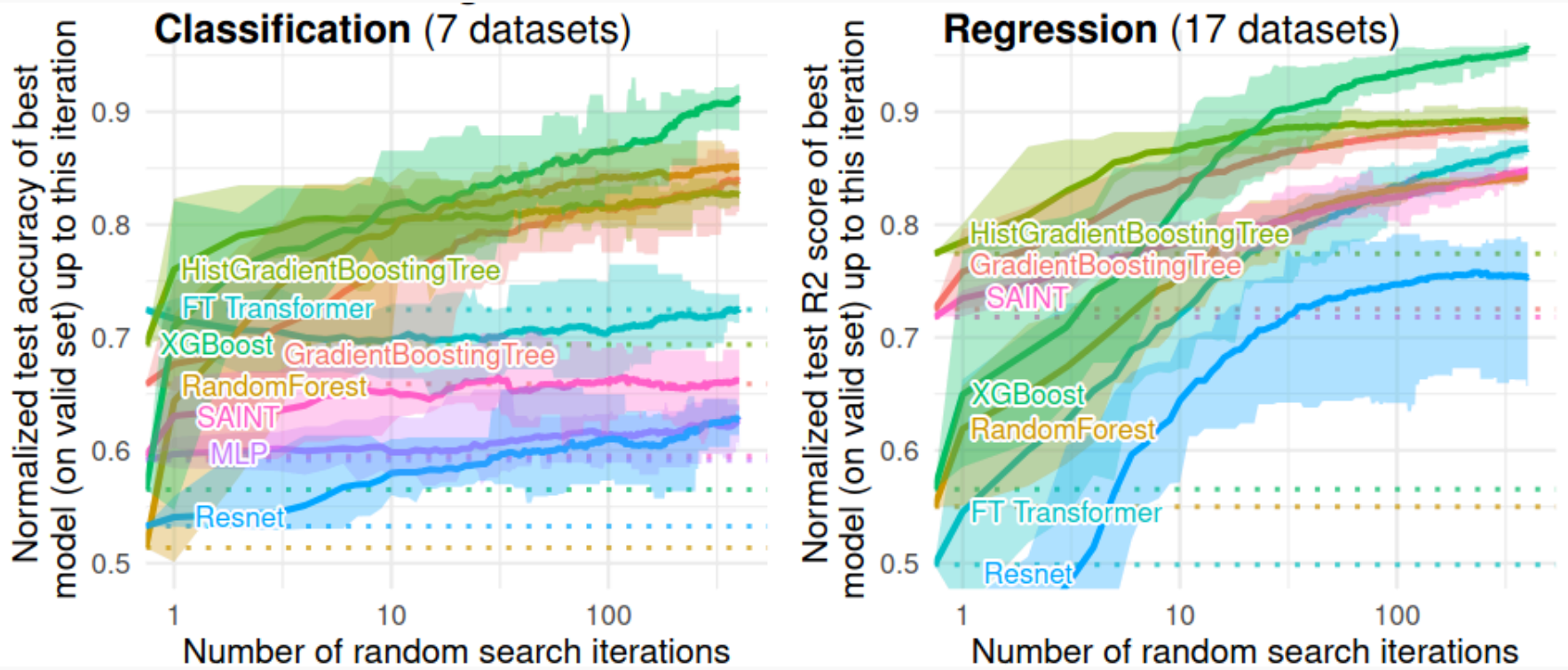


Typical dataset are mid-sized. This does not change with time.[1]

[1] https://www.kdnuggets.com/2020/07/poll-largest-dataset-analyzed-results.html

Deep learning underperforms on data tables

Tree-based methods outperform tailored deep learning architectures (Grinsztajn et al., 2022)



DAG for a RCT: the treatment is independent of the confounders

Other well known families of models

- Generalized linear models:
- Support vector machines:
- Gaussian processes:

Bibliography

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, 507–520.