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

Городилова Анастасия
Поляков Дмитрий
Шишков Алексей

Факультет Компьютерных Наук
НИУ ВШЭ

17 января 2023 г.

О чем статья?

- создание большого и качественного бенчмарка для табличных данных
- проведение развернутого сравнения классических ML моделей и нейросетей

Датасеты

- 45 датасетов выбрано по определенным критериям
- предобработка данных
- 2 режима исследования - medium и large

dataset_name	n_samples	n_features
credit	16714	10
california	20634	8
wine	2554	11
electricity	38474	7
covertype	566602	10
pol	10082	26
house_16H	13488	16
kdd_ipums_la_97-small	5188	20
MagicTelescope	13376	10
bank-marketing	10578	7
phoneme	3172	5
MiniBooNE	72998	50
Higgs	940160	24
eye_movements	7608	20
jannis	57580	54

Рис.: примеры собранных датасетов

Условия тестирования моделей

1. 400 итераций random-search делают 15 раз
2. 70% train, 9% val, 21% test
3. метрики
   - accuracy
   - r2
4. результаты по датасетам усредняются с помощью ADTM

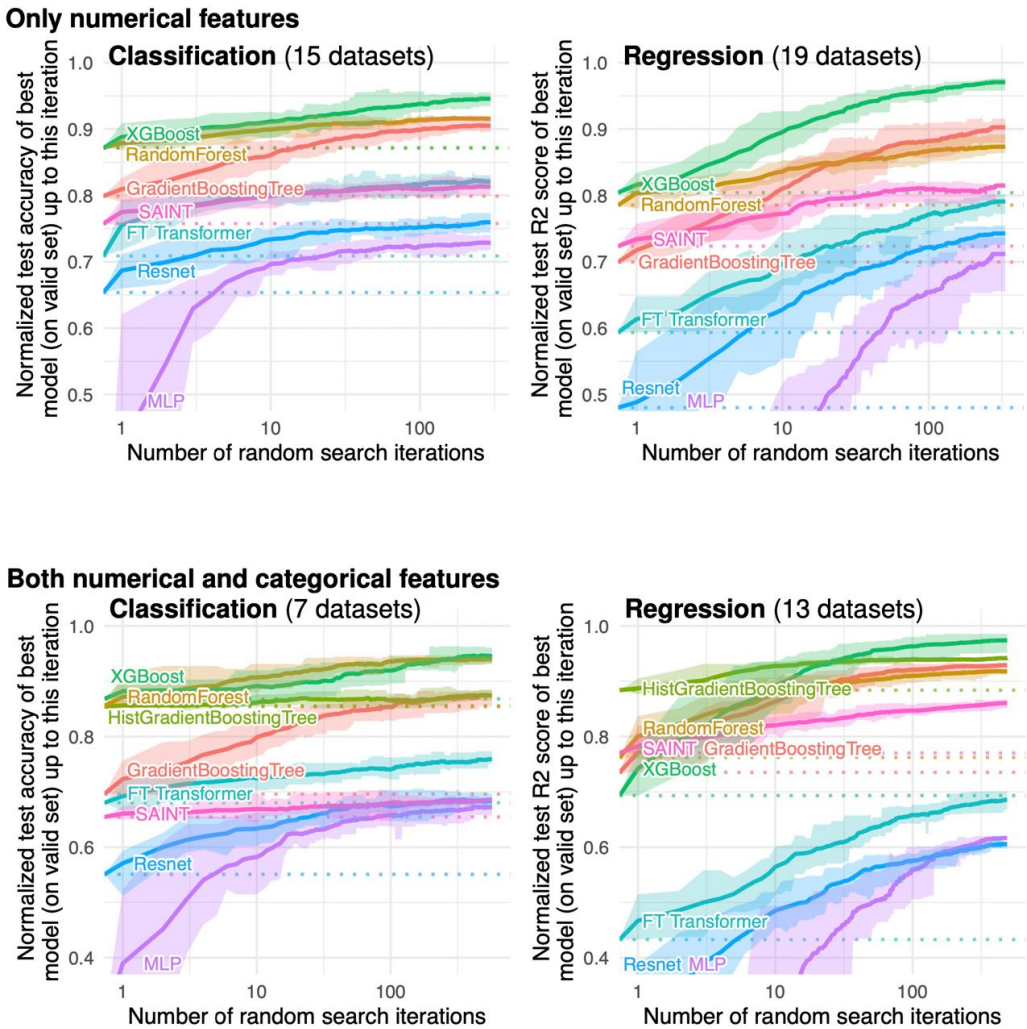
Используемые модели

ML модели:

- Random Forest
- GradientBoostingTrees
- XGBoost

DL модели:

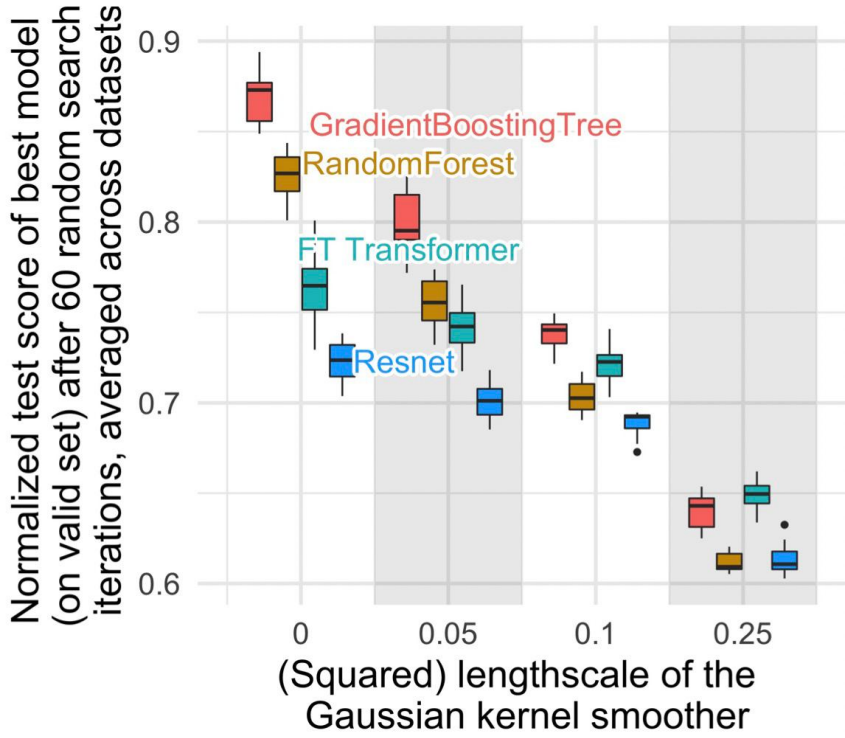
- MLP
- Resnet
- FT Transformer
- SAINT



Почему так происходит?

Гипотеза 1

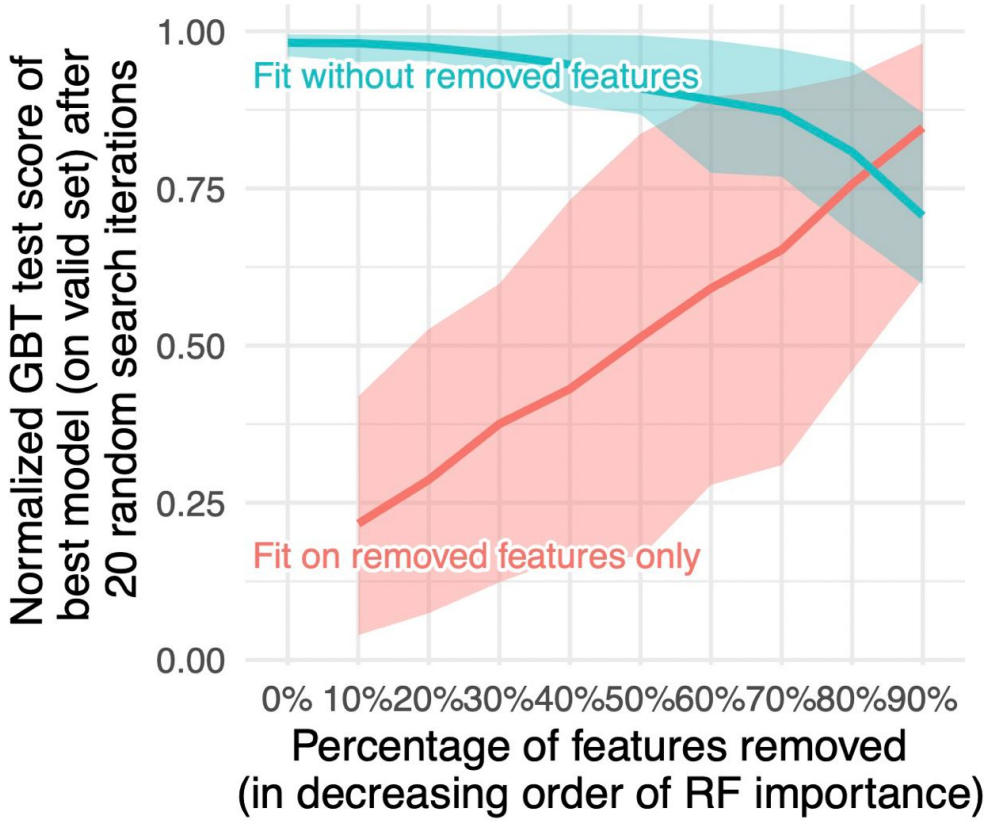
Гипотеза: нейросети предполагают, что данные гладкие;
Эксперимент: авторы сглаживают таргет Гауссовым фильтром;
Результат: качество ML моделей упало, нейросетей – не изменилось.



Почему так происходит?

Гипотеза 2

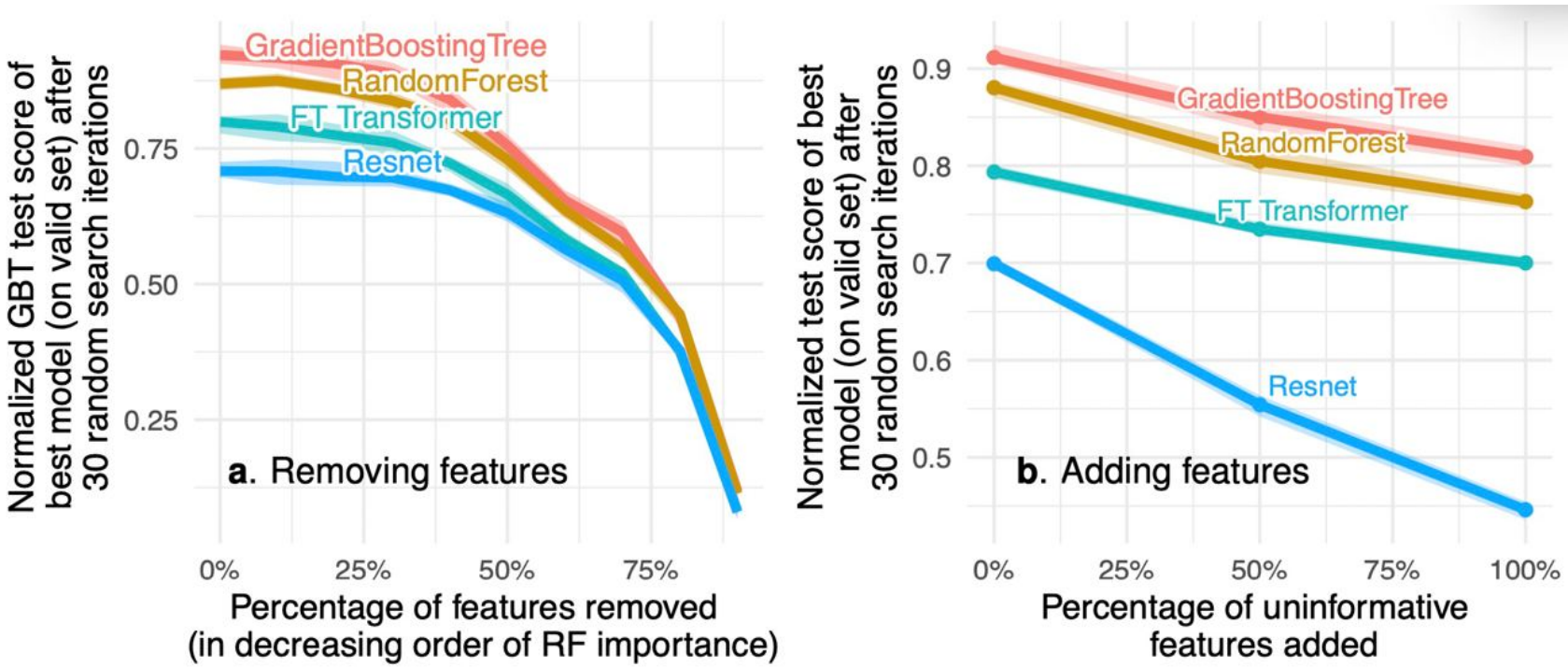
Гипотеза: влияние неинформативных фичей;
Эксперимент: удаление / добавление признаков;



Почему так происходит?

Гипотеза 2

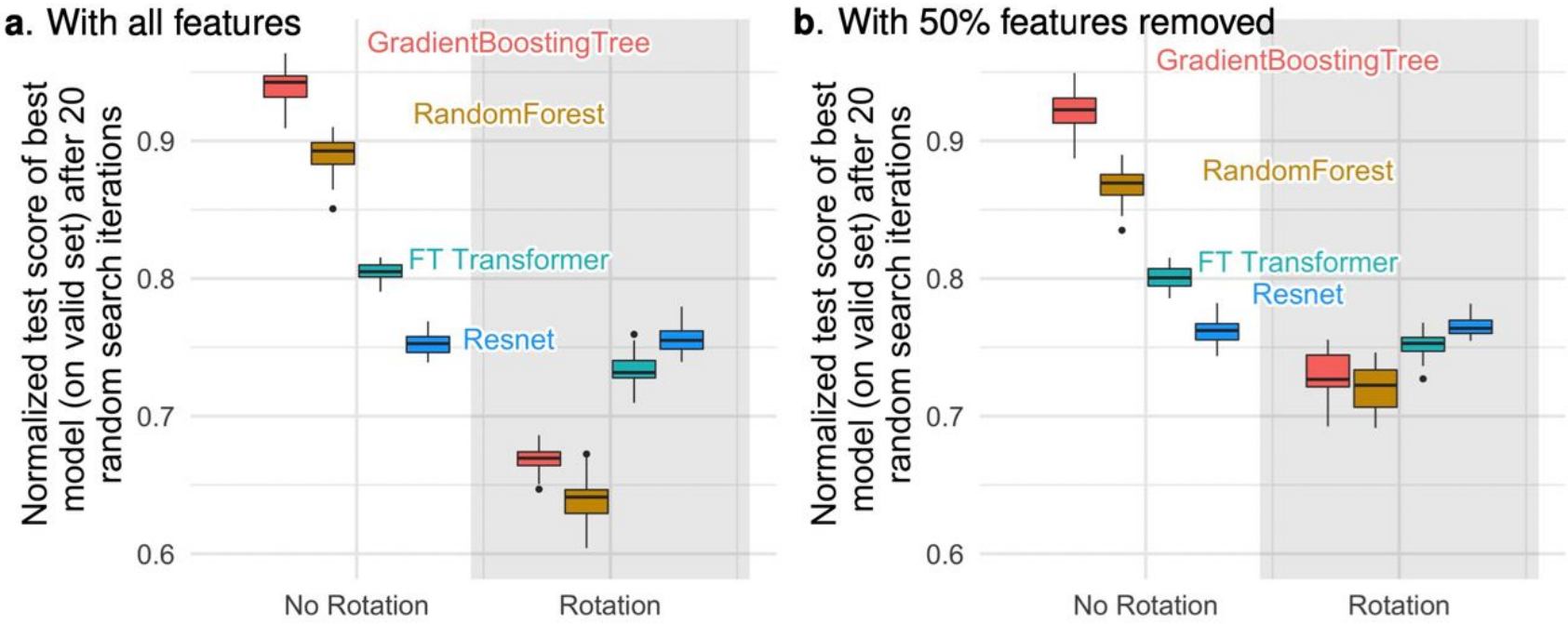
Результат: resnet очень сильно уязвим к неинформативным признакам.



Почему так происходит?

Гипотеза 3

Гипотеза: если алгоритм устойчив к повороту данных, то число данных для обучения будет линейно зависеть от количества плохих признаков;
Эксперимент: случайный поворот данных;
Результат: качество на resnet не меняется.



Вывод

Из-за чего проигрывают нейросети:

1. смещение в сторону сглаживания
2. неустойчивость к неинформативным признакам (resnet)
3. устойчивость к повороту данных