Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

2024 FS CAS PML - Supervised Learning
3 Regression
3.4 Modellvergleiche

Werner Dähler 2024

3 Regression - AGENDA

31. Einleitung
32. Regression klassisch (OLS)
33. Regression mit ML
34. Vergleiche über alle Modelle

3.4 Vergleiche über alle Modelle

- in einem integralen Vergleich werden alle hier bisher vorgestellten Methoden einander gegenübergestellt (ausser SVR und MLPRegressor), Tuningparameter werden, soweit aus den obigen Betrachtungen als relevant aufgefasst, hier so eingesetzt
  - `LinearRegression()`
  - `Ridge()`
  - `Lasso()`
  - `KNeighborsRegressor()`
  - `DecisionTreeRegressor(max_depth=8)`
  - `RandomForestRegressor()`
  - `AdaBoostRegressor(???)`, (vgl. Workshop 09)
  - `GradientBoostingRegressor()`
  - `HistGradientBoostingRegressor()`
  - `CatBoostRegressor(logging_level='Silent')`
  - `LGBMRegressor()`
- die untersten zwei sind optional, da nicht aus scikit-learn

3.4 Vergleiche über alle Modelle

- Metrik: R2 (r2_score)
- ausserdem wird noch für jedes Modell die Zeit gestoppt
- SVR und MLP werden aber wegen der obigen Erkenntnissen ausgeschlossen
- der ganze Vergleich wird in einem Loop abgearbeitet und die Teilresultate der einzelnen Modelle gesammelt und anschliessend visualisiert

- die einzelnen Schritte, analog Klassifikation (vgl. [ipynb])
  - importieren der Klassen
  - definieren und parametrisieren der Modelle
  - leere Listen zum Sammeln der Resultate der einzelnen Tests
    - names
    - scores
  - Loop über alle Modelle
  - visuelle Vergleiche

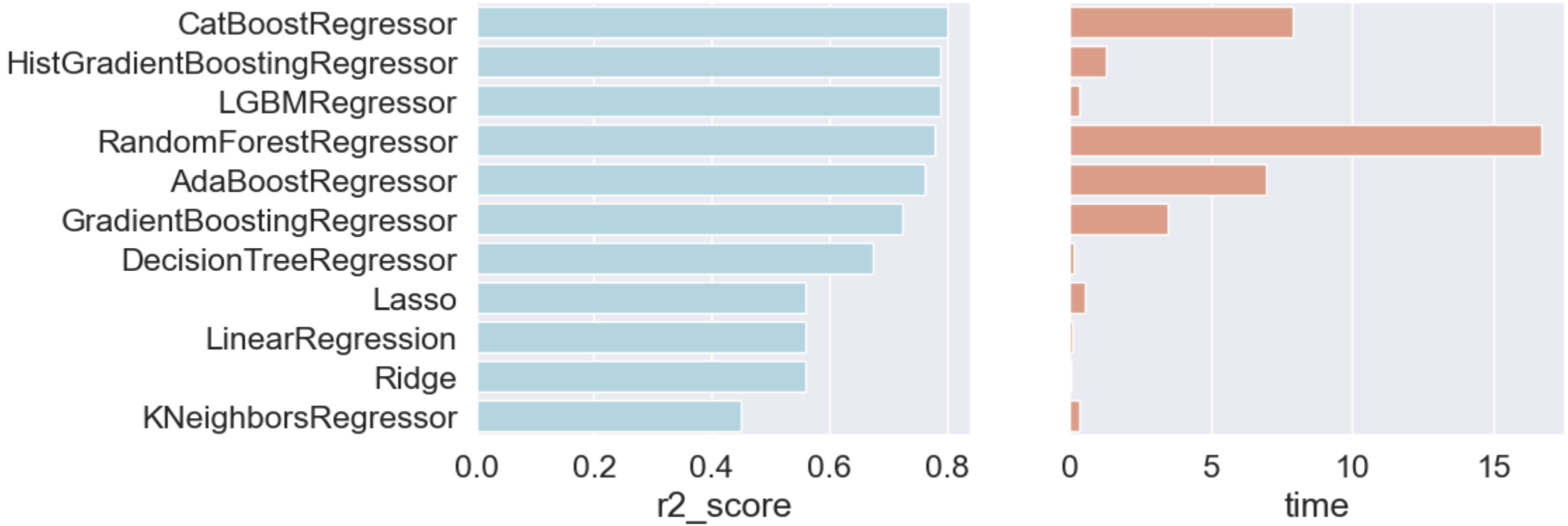
3.4 Vergleiche über alle Modelle

- Ergebnisse tabellarisch

Regressor	r2	time
LinearRegression	0.5601	0.092
Ridge	0.5601	0.019
Lasso	0.5601	0.550
KNeighborsRegressor	0.4493	0.334
DecisionTreeRegressor	0.6734	0.118
RandomForestRegressor	0.7779	16.676
AdaBoostRegressor	0.7616	6.929
GradientBoostingRegressor	0.7250	3.454
HistGradientBoostingRegressor	0.7891	1.300
CatBoostRegressor	0.8003	7.871
LGBMRegressor	0.7882	0.356

3.4 Vergleiche über alle Modelle

- Ergebnisse visuell



3.4 Vergleiche über alle Modelle

- bester Regressor mit Kennzahlen (vgl. [ipynb])

```
best_regressor : CatBoostRegressor
best_score     : 0.8003
best_time      : 7.8713
```

- und die Top 3 (vgl. [ipynb]):

	names	scores
8	CatBoostRegressor	0.800349
9	LGBMRegressor	0.788166
7	HistGradientBoostingRegressor	0.785776

3.4 Vergleiche über alle Modelle

Fazit:

- LinearRegression, Lasso und Ridge weisen vergleichbare Performance auf und sind auch zeitlich gut unterwegs, wie erwartet bringen Lasso und Ridge bei der vorliegenden Datenlage nur marginale oder keine Verbesserung gegenüber LinearRegression
- KNeighborsRegressor hat die schlechteste Performance (Standard Parametrisierung)
- beste Performance finden wir bei CatBoostRegressor (r2: 0.8003), gefolgt von HistGradientBoostingRegressor und LGBMRegressor

Vorschlag für weiteres Vorgehen:

- die drei genannten Regressoren mittels Parameter Tuning weiter optimieren

3.4 Vergleiche über alle Modelle

Workshop 10

Gruppen zu 2 bis 4, Zeit: 30'

- Vergleichen Sie alle Regressoren (ausser `SVR` und `MLPRegressor`) mit folgenden Modifikationen
  - die Vergleiche werden ohne und mit Standardisierung der Features durchgeführt
  - die Resultate (r2) werden in Form einer Heatmap zusammengestellt

- informieren Sie sich zum Vorgehen am Code in `3.4 Regression - Modellvergleiche.ipynb`
- Präsentation der Ergebnisse als
  - seaborn heatmap
  - alternative Visualisierung: Grouped barplots