

Python - projekt - prezentacja

2018-07-01

Kurs Junior Data Scientist Zaoczne 1 (JDSZ1)

Raczki

Bartosz, Filip, Monika Kucal, Piotr

- 1. Dane**
- 2. Machine Learning - GridSearchCV**
- 3. Machine Learning - Personal Best Models**

1. Dane - Eksploracja, normalizacja, skalowanie, selekcja cech i obserwacji

Dane

- Zbiór danych Kaggle: [Rowery](#)
- Cel: Prognoza liczby wypożyczonych rowerów
 - **datetime** - hourly date + timestamp
 - **season** - 1 = spring, 2 = summer, 3 = fall, 4 = winter
 - **holiday** - whether the day is considered a holiday
 - **workingday** - whether the day is neither a weekend nor holiday
 - **weather** - 1: Clear, Few clouds, Partly cloudy, Partly cloudy; 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist; 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds; 4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog
 - **temp** - temperature in Celsius
 - **atemp** - "feels like" temperature in Celsius
 - **humidity** - relative humidity
 - **windspeed** - wind speed
 - **casual** - number of non-registered user rentals initiated
 - **registered** - number of registered user rentals initiated
 - **count** - number of total rentals

Dane

	MK	FJ	BG	PS
Wczytanie	pandas.read_csv	pandas.read_csv	pandas.read_csv	pandas.read_csv
Eksploracja	korelacje, wykresy, rozkłady	korelacje, wykresy, rozkłady	korelacje, wykresy, rozkłady	Gęstość, rozkłady, korelacje, skośność
Normalizacja	-	-	-	-
Skalowanie	+	-	-	-
Selekcja	eliminacja: atemp, day	eliminacja : atemp, day	eliminacja : atemp, day	Eliminacja: atemp, season
Inne	y1: casual y2: registered	y1: casual y2: registered	y1: casual y2: registered	Y: count, podział daty na więcej cech, eliminacja zer w wilgotności, one-hot encoding
Train/Test	80/20 (seed 789)	80/20 (seed 789)	80/20 (seed 789)	80/20 (seed 789)

2. Machine Learning - GridSearchCV

Machine Learning - GridSearchCV

	MK	FJ	BG	PS
Model	Decision Tree	RF - zmiana na KKN	SVR	XGBoost
Hiperparametry	y1: casual criterion: mse max depth: 13 min samples split: 10 min samples leaf: 10 y2: registered criterion: mse max depth: 14 min samples split: 20 min samples leaf: 10	y1: casual leaf_size=100 n_neighbors=10 P: 2 metric='minkowski' y2: registered leaf_size=30 n_neighbors=5 P: 2 metric='minkowski'	y1: casual C=0.5 tol=0.1 dual=True epsilon=0.1 loss='squared_epsilon_insensitive' y2: registered C=1 tol=0.1 dual=True epsilon=0.01 loss='squared_epsilon_insensitive'	Y: count colsample_bylevel: 0.8 colsample_bytree: 0.8 max_depth: 8, min_child_weight: 3 n_estimators: 200
RMLSE - Test	0.35	0.88	1. 45	0.35
RMLSE - Kaggle	0.49	0.96	1.45	0.58

3. Machine Learning - Personal Best Models

|



Dziękujemy!

Pytania?
Slack / email