



PREDICTIVE POWER SCORE PYTHONBAN

Szabó Dóra, PhD
Data Scientist, Spicy Analytics

2021. ÁPRILIS 15.
BUDAPEST

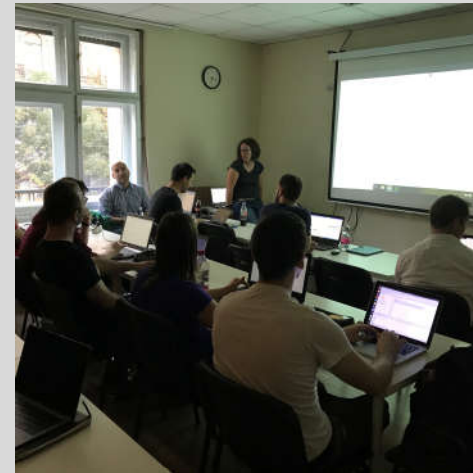
ML FORUM

ONLINE

MACHINE LEARNING | DATA SCIENCE | ARTIFICIAL INTELLIGENCE

Magamról

- Biológia, tudományos kutatás
- Etológia PhD
 - Viselkedés és kognitív idegtudomány
- Adatelemzés vs. programozás



Az első BudapestPy Workshop

Első adatos önképzőkör

DS Spicy Analysis

I Budapest ML Forum

2019 január

2019 szeptember

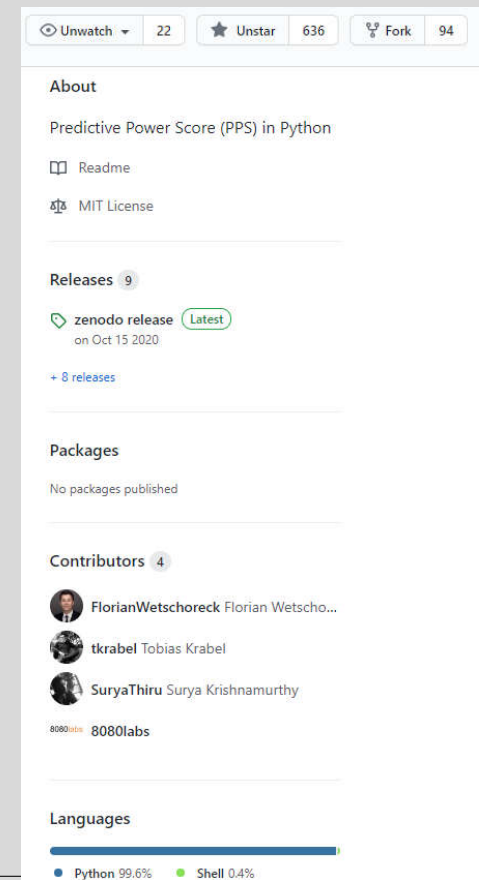
2020 február

COVID-19

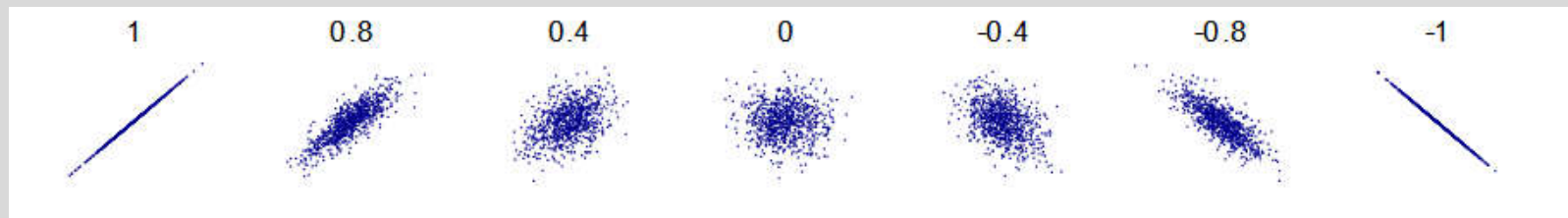
2021 április

A Predictive Power Score csomag

- Pythonban írva (open source)
- [Részletes dokumentáció](#), [egyszerű használat](#)
- Felgyorsítja az adatmegértést
- Változók közti kapcsolatok feltérképezése
- Egyszerű, gyors vizualizáció
- A korrelációs mátrix egy lehetséges alternatívája
- Egy lehetséges eszköz feature selection során



A korrelációs mátrix határai



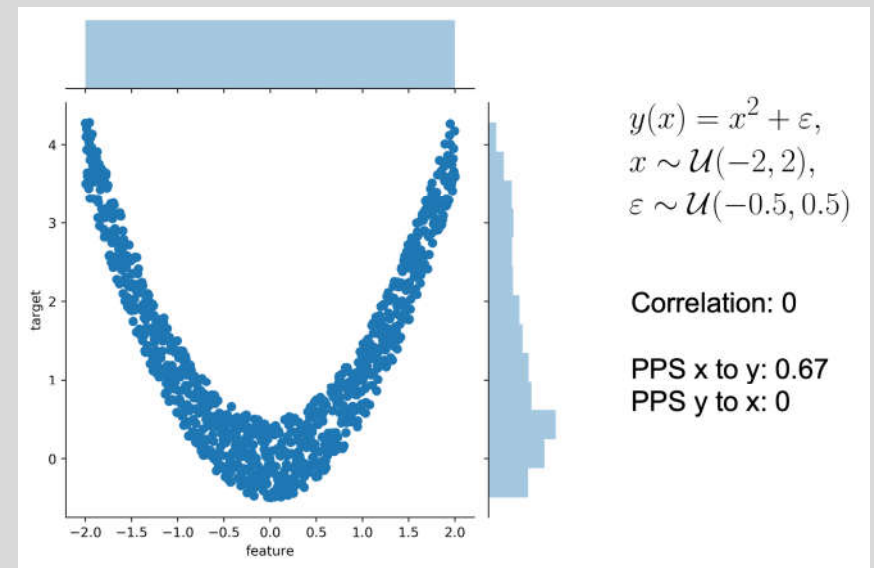
- Folytonos változók → **Pearson** (lineáris)
- Kategórikus/Ordinális → **Spearman's** (monoton)
- A kapcsolatot **erőssége és iránya**
- **Szimmetrikus** korrelációs mátrix
- Számos mintázatot nem jelez
- Bináris/Nominális változók **kihagyása**

Predictive Power Score

**„an asymmetric, data-type-agnostic score
for predictive relationships between two columns
that ranges from 0 to 1”**

Miben tud többet a PPS?

- **Nominális és bináris változókat** is kezel
- **Aszimmetrikus** kapcsolat a változók közt
- Könnyebben észrevehető **potential information leakage**
- Értelmezhető **irányított gráfként**
- „okozati” láncok feltérképezése
- **Out-of-box kezel/jelez:**
 - id és konstans változók



Mit is számol?

Célváltozó	Algoritmus	Kiértékelés	Alsó limit (naiv predikció)	Felső limit	PPS
Numerikus	Decision Tree Regressor	Mean Absolute Error (MAE)	Medián érték	0	$1 - (\text{model_mae} / \text{naive_mae})^*$
Kategorikus, bináris	Decision Tree Classifier	weighted F1	Leggyakoribb kategória	1	$(\text{model_f1} - \text{baseline_f1}) / (1.0 - \text{baseline_f1})$

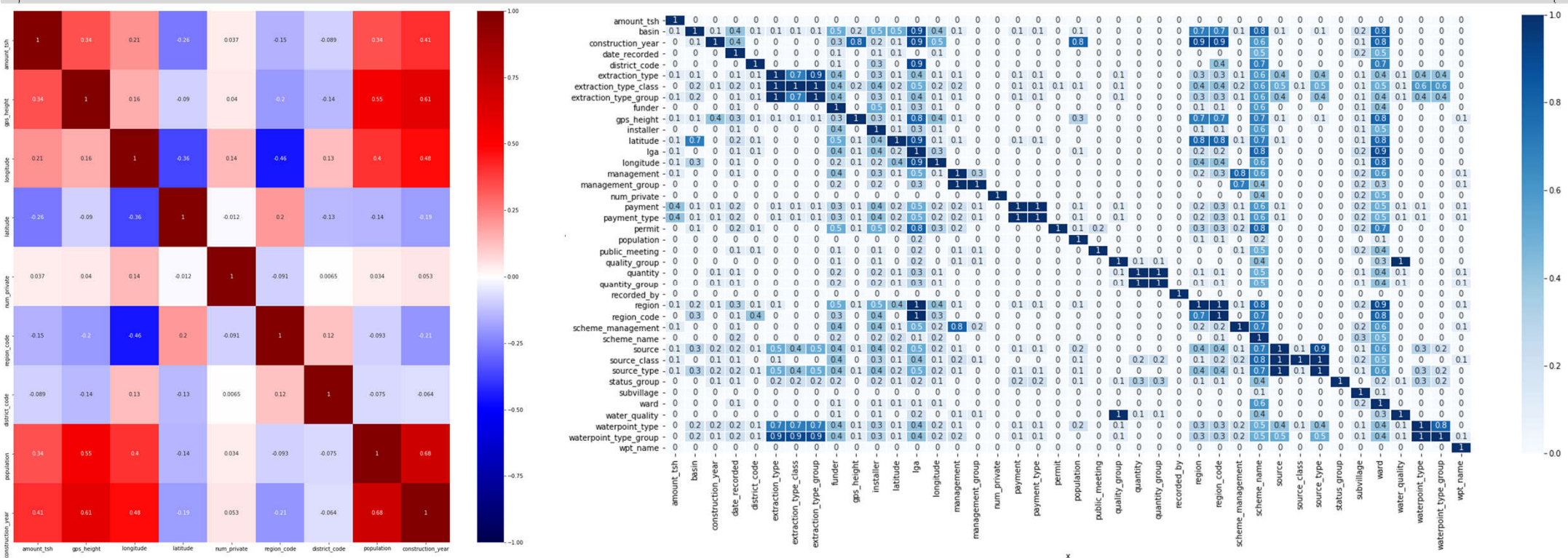
Döntési fa (alapértelmezettként 4-fold CV a the test set-en)

A teljesítményt a naiv/baseline modellhez viszonyítja

Pl. $(0.95 - 0.1) / (1 - 0.1) = 0.94$

*If $\text{model_mae} > \text{naive_mae}$
PPS=0

Korrelációs mátrix vs. PPS mátrix



9*9

40*40

A csomag 3 valódi funkciója

❖ `pps.score(df, "x", "y")`

❖ `pps.predictors(df, y="y")`

❖ `pps.matrix(df)`

A csomag 3 valódi funkciója

❖ `pps.score(df, "x", "y")`

❖ `pps.predictors(df, y="y")`

❖ `pps.matrix(df)`

```
pps.score(pp_df, "latitude", "region")
```

```
{'x': 'latitude',  
'y': 'region',  
'ppscore': 0.3694664457814855,  
'case': 'classification',  
'is_valid_score': True,  
'metric': 'weighted F1',  
'baseline_score': 0.0634,  
'model_score': 0.4094422731189393,  
'model': DecisionTreeClassifier()}
```

```
pps.score(pp_df, "region", "latitude" )
```

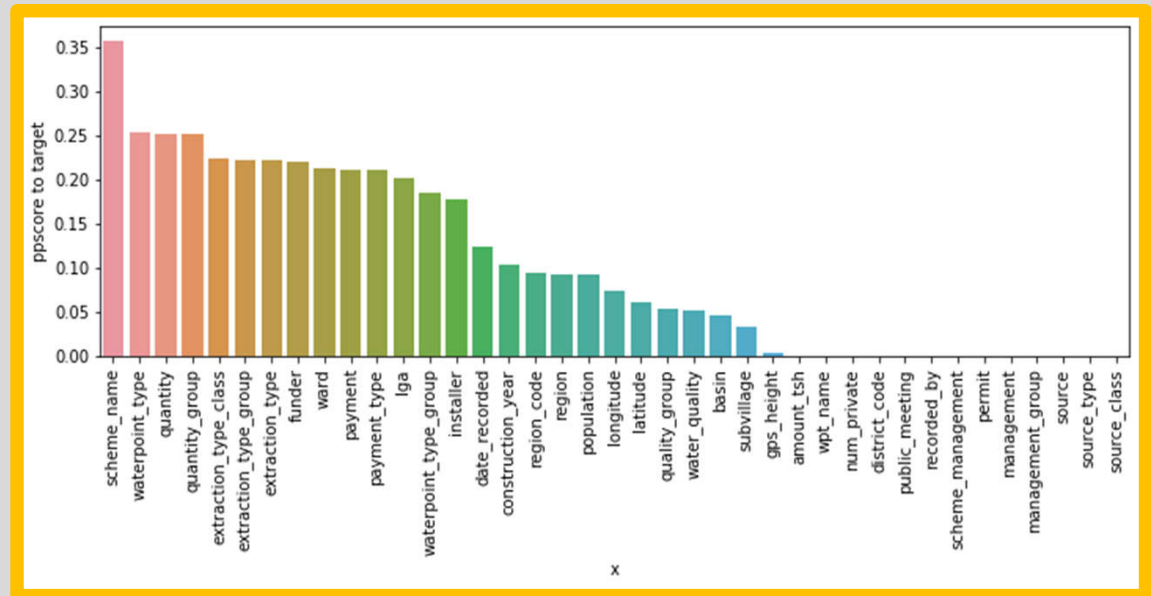
```
{'x': 'region',  
'y': 'latitude',  
'ppscore': 0.8008287806264682,  
'case': 'regression',  
'is_valid_score': True,  
'metric': 'mean absolute error',  
'baseline_score': 2.5551732796919997,  
'model_score': 0.5089169778269218,  
'model': DecisionTreeRegressor()}
```

A csomag 3 valódi funkciója

❖ `pps.score(df, "x", "y")`

❖ `pps.predictors(df, y="y")`

❖ `pps.matrix(df)`



```
import ppscore as pp
predictors_df = pp.predictors(pp_df, y="status_group")
```

```
predictors_df.head()
```

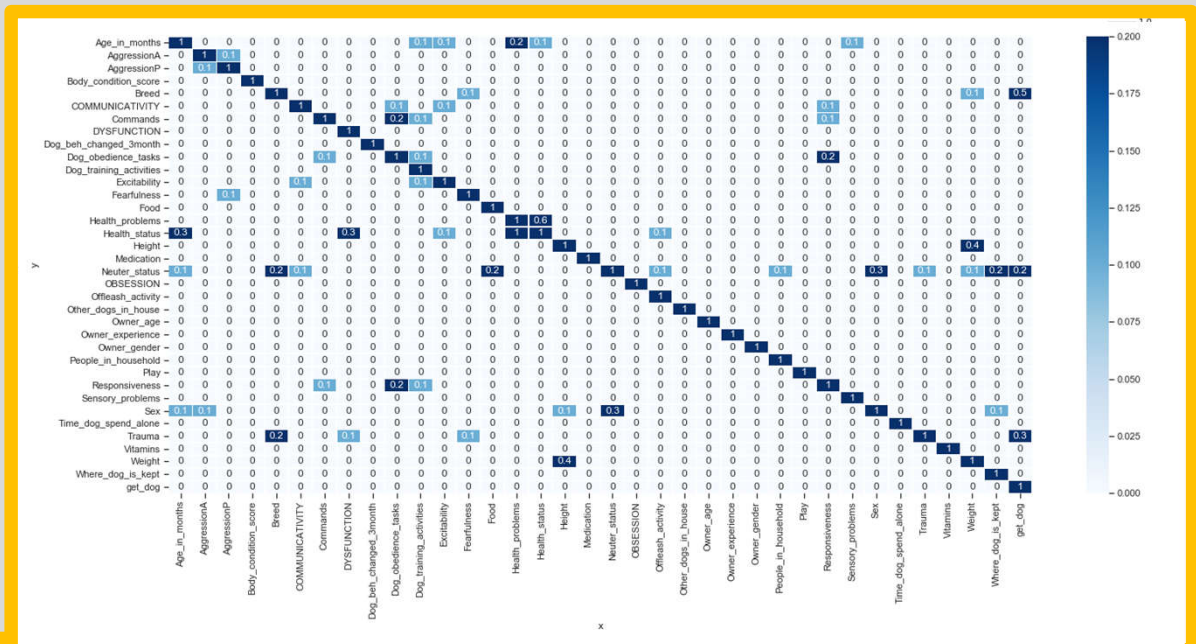
	x	y	ppscore	case	is_valid_score	metric	baseline_score	model_score	model
0	scheme_name	status_group	0.355831	classification	True	weighted F1	0.4448	0.642357	DecisionTreeClassifier()
1	waterpoint_type	status_group	0.254172	classification	True	weighted F1	0.4428	0.584425	DecisionTreeClassifier()
2	quantity	status_group	0.251796	classification	True	weighted F1	0.4428	0.583101	DecisionTreeClassifier()
3	quantity_group	status_group	0.251796	classification	True	weighted F1	0.4428	0.583101	DecisionTreeClassifier()
4	extraction_type_class	status_group	0.223891	classification	True	weighted F1	0.4428	0.567552	DecisionTreeClassifier()

A csomag 3 valódi funkciója

❖ `pps.score(df, "x", "y")`

❖ `pps.predictors(df, y="y")`

❖ `pps.matrix(df)`



```
import ppscore as pp
predictors_df = pp.predictors(pp_df, y="status_group")
```

```
predictors_df.head()
```

	x	y	ppscore	case	is_valid_score	metric	baseline_score	model_score	model
0	scheme_name	status_group	0.355831	classification	True	weighted F1	0.4448	0.642357	DecisionTreeClassifier()
1	waterpoint_type	status_group	0.254172	classification	True	weighted F1	0.4428	0.584425	DecisionTreeClassifier()
2	quantity	status_group	0.251796	classification	True	weighted F1	0.4428	0.583101	DecisionTreeClassifier()
3	quantity_group	status_group	0.251796	classification	True	weighted F1	0.4428	0.583101	DecisionTreeClassifier()
4	extraction_type_class	status_group	0.223891	classification	True	weighted F1	0.4428	0.567552	DecisionTreeClassifier()

A PPS leegyszerűsített értelmezése

- PPS: **hány százalékos teljesítménynövekedést** jelent?
 - naiv modell vs. 1 változós modell (ideális, determinisztikus modellhez viszonyítva)
- **Kontextusfüggő** (pl. a teljes mátrix értékeihez képest)
- **Irányadó értékek a PPS értelmezéséhez:**
 - $PPS = 0$ **nem** tartalmaz a predikció szempontjából **releváns információt**
 - $0 < PPS \leq 0.2$ releváns, **de gyenge**
 - $0.2 < PPS \leq 0.8$ releváns, **jelentős**
 - $0.8 < PPS$ túl szép, pl. **determinisztikus kapcsolatra utalhat**

A hátrányok

Korrelációs mátrixhoz képest:

- Lassabb (40 oszlop esetén 1-10 perc között)
- Nehezebben/bonyolultabban kommunikálható (nemlineáris kapcsolatok is)
- Target specifikus: a különböző target változókhoz számolt PPS-k szigorúan véve nem összemérhetőek (különböző baseline-ok)

A hátrányok

Univerzális:

- Nem tud interakciókat kezelni (vö. feature importance)
- A sampling/mintavételezés nem biztos, hogy megfelelő idősoros adatok esetén (stratifiedKFold, normal Kfold)
- GPU jelenleg nem támogatott
- NEM váltja ki a forward/backward selection lépést a modell szelekció során!

Összességében

- Egyszerű, könnyen használható
 - Trade off: limitált performance metrikák
 - Automatikusan felismeri a változótípusokat
 - DE minden numerikusnál regresszió az alapértelmezett
 - Gyors áttekintést tud nyújtani
 - Segíthet új összefüggéseket, láncolatokat észlelni
 - Nem jelzi a két változó közötti kapcsolat típusát
 - többet kell dolgozni vele
 - Nagy információsűrűség, nehezebben emészthető
 - több magyarázatot, segítséget igényelhet az interpretációja
- +1 hasznos tool, nem silver bullet

Köszönöm a figyelmet!