ZMUM – Projekt 1

Piotr Olesiejuk

Wstępny preprocessing

- Usunięcie kolumn o ponad 95% NA
- Zamiana zmiennych liczbowych o mniej niż 10 poziomach na factory
- W zmiennych kategorycznych uzupełnienie:

    a) pustych wartości jako : „Empty”

    b) NA jako : „Nas”

Dalsze dwa podejścia

1. Usuwanie kolumn kategorycznych na podstawie pewnej bramki
2. Indywidualne przekształcanie zmiennych

Indywidualne przekształcenia

	0	1	1_perc	0_perc	all_perc	0_perc_from_all_zeros	1_perc_from_all_ones
7	17272	1589	0.08424792	0.9157521	0.471525	4.659419e-01	0.54213579
0	9845	621	0.05933499	0.9406650	0.261650	2.655858e-01	0.21187308
Nas	4233	176	0.03991835	0.9600817	0.110225	1.141925e-01	0.06004777
14	3683	337	0.08383085	0.9161692	0.100500	9.935526e-02	0.11497782
21	1326	136	0.09302326	0.9069767	0.036550	3.577113e-02	0.04640055
28	524	52	0.09027778	0.9097222	0.014400	1.413580e-02	0.01774139
35	184	20	0.09803922	0.9019608	0.005100	4.963716e-03	0.00682361
140	1	0	0.00000000	1.0000000	0.000025	2.697672e-05	0.00000000
42	1	0	0.00000000	1.0000000	0.000025	2.697672e-05	0.00000000

	0	1	1_perc	0_perc	all_perc	0_perc_from_all_zeros	1_perc_from_all_ones
7	17272	1589	0.08424792	0.9157521	0.471525	0.4659419	0.54213579
0	9845	621	0.05933499	0.9406650	0.261650	0.2655858	0.21187308
Other	5719	545	0.08700511	0.9129949	0.156600	0.1542799	0.18594336
Nas	4233	176	0.03991835	0.9600817	0.110225	0.1141925	0.06004777

Normalizacja zmiennych numerycznych

- Wewnątrz kroswalidacji
- Wypełnianie brakujących wartości losowo:
    a) medianą
    b) średnią uciętą z losowym ucięciem z przedziału (0.01 , 0.03)
- Skalowanie do przedziału (0,1)

Zastosowane algorytmy

- randomForest (ntree = 60)
- rpart (cp = 0.01)
- xgboost (nrounds = 10, eta = 0.04, b_score = 0.92)
- kNN

Wyniki

Algorytm	**Precyzja 10%**
randomForest	39,08%* (σ = 0,15%)
xgboost	38,92 %** (σ = 0,22%)
rpart	37,69%** (σ = 0,23%)
kNN	-----

*Precyzja 10% obliczona jako średnia z wyników 5-krotnie powtórzonej dziesięciokrotnej kroswalidacji
**Precyzja 10% obliczona jako średnia z wyników 10-krotnie powtórzonej dziesięciokrotnej kroswalidacji

kNN

	fit	
test_class	0	Row Total
0	7412	7412
	0.926	
1	588	588
	0.073	
Column Total	8000	8000

Rekomendowany: randomForest

- Najwyższy wynik precyzji
- Mniejsza wariancja precyzji po kroswalidacji niż w xgboost

Dziękuję za uwagę