Praca Domowa 2

Julia Girtler

2023-11-20

Wstęp

Przed budowaniem naszych modeli należało odpowiednio przygotować dane. Kolumny zmiennych kategorycznych zostały przekształcone na zestaw nowych kolumn binarnych. Zastosowane zostało drop_first = True, aby uniknąć tzw. “pułapki zmiennej zerojedynkowej” (dwie kolumny zależne od siebie).

Następnie podzieliśmy dane na zestaw trenigowy i testowy z parametrem train_split = 0.15 (niski, ponieważ mało obserwacji).

Część 1

Dla każdego z typów regularyzacji został przeprowadzony Gridsearch z kroswalidacją pięciostopniową. Następnie dla uzyskanych najlepszych parametrów zostały policzone miary dla zbioru treningowego i testowego: accuracy, preccision, recall, f1_score, roc_auc_score.

Model regresji logistycznej: penalty = ‘none’, solver =‘lbfgs’

miara	treningowy	testowy
accuracy	0.778	0.753
recall	0.887	0.861
precission	0.811	0.809
f1 score	0.847	0.834
roc_auc score	0.831	0.753

Model regresji logistycznej z regularyzacją L1: penalty=“l1”, C=1, solver=‘liblinear’

miara	treningowy	testowy
accuracy	0.779	0.760
recall	0.892	0.861
precission	0.810	0.816
f1 score	0.849	0.838
roc_auc score	0.839	0.750

Model regresji logistycznej z regularyzacją L2: penalty = “l2”, C=0.01, solver=‘lbfgs’

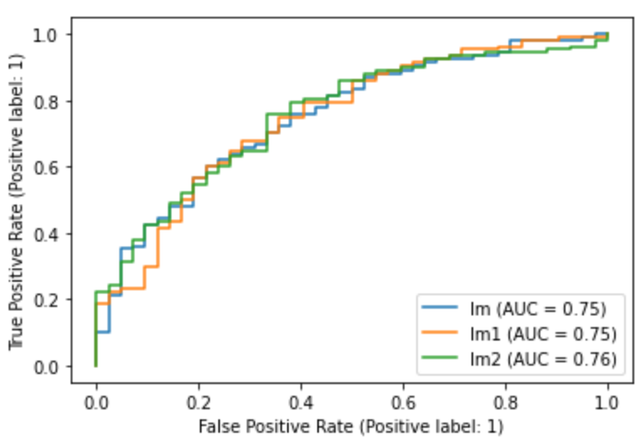
miara	treningowy	testowy
accuracy	0.724	0.720
recall	0.954	0.954
precission	0.731	0.736
f1 score	0.828	0.831
roc_auc score	0.763	0.757

Porównanie modeli

Zostały porównane następujące modele: model regresji logistycznej bez regularyzacji, model regrasji logistycznej z regularyzacją L1 oraz model regresji logistycznej z regularyzajcą L2

miara	lm	lm1	lm2
accuracy	0.753	0.760	0.720
recall	0.861	0.861	0.954
precission	0.809	0.816	0.736
f1 score	0.834	0.838	0.831
roc_auc score	0.753	0.750	0.757

Krzywe ROC



Wnioski

- najlepsze accuracy, preccision, f1_score otrzymaliśmy dla modelu z regularyzacją L1, natomiast recall i roc_auc_score dla modelu z regularyzacją L2
- możemy odczytać z wykresu, że najlepsze AUC otrzymaliśmy dla modelu regresji logistycznej z regularyzacją L2
- model bez regularyzacji wypadł najgorzej
- Obliczone zostały współczyniki dla modelu regresji logistycznej z regularyzacją L1. Niektóre z nich były równe 0, co oznacza brak wpływu na podejmowane przez model decyzje. Były to następujące zmienne:

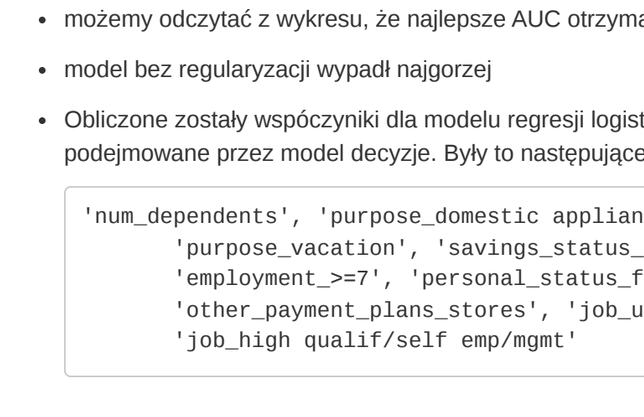
```
'num_dependents', 'purpose_domestic appliance', 'purpose_repairs',
       'purpose_vacation', 'savings_status_100<=X<500', 'employment_1<=X<4',
       'employment_>=7', 'personal_status_female single',
       'other_payment_plans_stores', 'job_unskilled resident',
       'job_high qualif/self emp/mgmt'
```

Część 2

W tej częsci został stworzony model oparty na metodzie wektorów podpierających (support vectotr machine).

Powstał nowy model X_new, w którym zostały usunięte kolumny wymienione wyżej we wnioskach części 1.

miara	treningowy.nowy	testowy.nowy	treningowy.stary	testowy.stary
accuracy	0.766	0.760	0.767	0.767
recall	0.861	0.852	0.861	0.843
precission	0.813	0.821	0.815	0.835
f1 score	1.000	0.836	0.837	0.839



Wnioski:

- dla nowego zestawu danych dla zbioru testowego lepsze okazało się recall, jednak reszta miar jest wyższa dla starych danych(bez redukcji kolumn), może to wynikać z regularyzacji modelu