Praca domowa 2 - Wstęp do uczenia maszynowego

Oliwia Trzcińska

1 Wstęp

Celem pracy domowej jest zbadanie, jak działają modele regresji logistycznej (z regularyzacją oraz bez) oraz maszyny wektorów podpierających w problemie klasyfikacji. Do tego celu będziemy używać rzeczywistego zbioru danych `credit-g`.

2 Dane

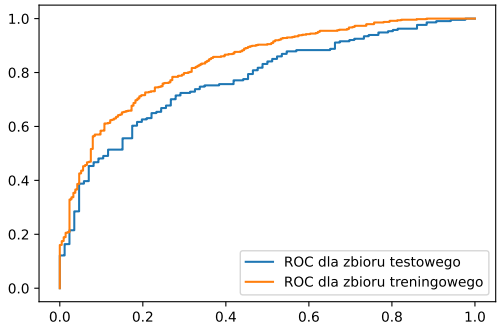
Po pobraniu danych musimy przygotować je do dalszej pracy. Ponieważ istnieją kolumny ze zmiennymi nominalnymi, zamieniamy je na zmienne binarne. Dzielimy dane na zbiory treningowy i testowy w proporcji 7:3, ustawiając wartość parametru `random_state = 320584`.

3 Modele regresji logistycznej

3.1 Model regresji logistycznej bez regularyzacji

Model regresji logistycznej tworzymy bez dopasowywania jego parametrów. Wyznaczamy miary tego modelu na zbiorach treningowym i testowym. Otrzymujemy wyniki przedstawione w tabeli. Narysujemy także krzywe ROC.

Miara	Wartość na zbiorze treningowym	Wartość na zbiorze testowym
dokładność	0,786	0.723
czułość	0,891	0.794
precyzja	0,817	0,813
wartość AUC	0,838	0,774

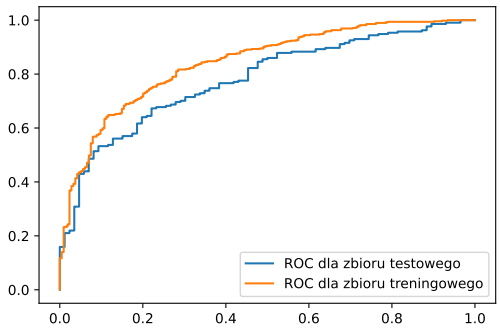


3.2 Model regresji logistycznej z regularyzacją L1

Do stworzenia modelu regresji logistycznej z regularyzacją L1 znajdziemy optymalny parametr `C`. Użyjemy 5-krotnej walidacji, zbadamy kilka wartości `C` i wybierzemy optymalną. W naszym przypadku

chcemy unikać wartości `false positive`, ponieważ zaklasyfikowanie złego klienta jako dobrego może mieć potencjalnie gorsze konsekwencje niż zaklasyfikowanie dobrego jako złego. Dlatego wybierzemy takie `C`, które maksymalizuje precyzję - w naszym przypadku będzie ono równe 0,99.
Dla modelu z najlepszym parametrem otrzymaliśmy miary przedstawione w tabeli. Stworzymy także krzywe ROC.

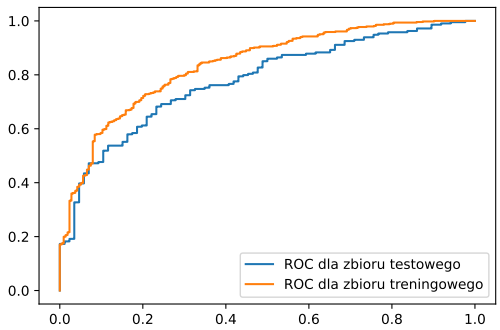
Miara	Wartość na zbiorze treningowym	Wartość na zbiorze testowym
dokładność	0,783	0.74
czułość	0,891	0.822
precyzja	0,814	0,815
wartość AUC	0,843	0,781



3.3 Model regresji logistycznej z regularyzacją L2

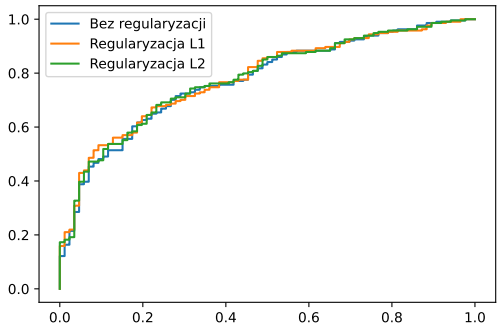
Model regresji logistycznej z regularyzacją L2 tworzymy analogicznie jak poprzednio. Tym razem optymalną wartością parametru jest `C = 0,5`. Dla modelu z tym `C` miary są przedstawione w tabeli, poniżej znajdują się także krzywe ROC.

Miara	Wartość na zbiorze treningowym	Wartość na zbiorze testowym
dokładność	0,79	0.727
czułość	0,899	0.808
precyzja	0,817	0,808
wartość AUC	0,838	0,779



3.4 Interpretacja wyników

Aby móc wygodniej porównać, jak z problemem klasyfikacji poradziły sobie wyżej opisane modele, przedstawimy na jednym wykresie krzywe ROC dla zbioru testowego dla różnych modeli. Okazuje się, że regularyzacja lub jej brak w naszym przypadku nie wpływa znacząco na jakość modelu.



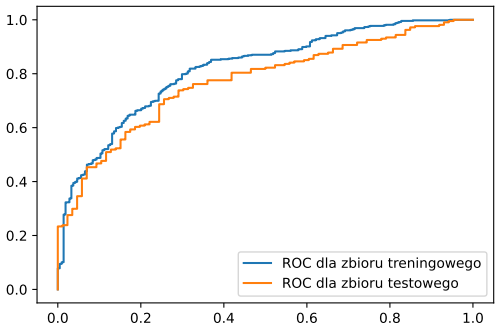
Sprawdzimy teraz, które zmienne są nieistotne w każdym ze stworzonych przez nas modeli. Zmienną uważamy za niestotną, kiedy wartość odpowiadającego jej współczynnika w modelu jest równa 0. Są to odpowiednio:

- dla modelu bez regularyzacji: `purpose_vacation`, `personal_status_female single`,
- dla modelu z regularyzacją L1: `checking_status_0<=X<200`, `purpose_furniture/equipment`, `purpose_domestic appliance`, `purpose_vacation`, `purpose_retraining`, `purpose_business`, `savings_status_100<=X<500`, `savings_status_500<=X<1000`, `employment_unemployed`, `employment_>=7`, `personal_status_male div/sep`, `personal_status_female single`, `other_parties_none`, `other_parties_co applicant`, `property_magnitude_life insurance`, `other_payment_plans_bank`, `other_payment_plans_stores`, `job_unskilled resident`, `job_skilled`, `own_telephone_none`, `foreign_worker_yes`,
- dla modelu z regularyzacją L2: `purpose_vacation`, `personal_status_female single`.

4 Model wektorów podpierających (SVM)

Stworzymy teraz model wektorów podpierających. Na wejściu ograniczamy liczbę zmiennych - usuwamy te, które były nieistotne dla modelu z regularyzacją L1. Ze względu na zbyt długi czas obliczeń, nie będziemy optymalizować parametru `C`. Dla tego modelu miary są przedstawione w tabeli, poniżej znajdują się także krzywe ROC.

Miara	Wartość na zbiorze treningowym	Wartość na zbiorze testowym
dokładność	0,78	0.733
czułość	0,852	0.762
precyzja	0,835	0,849
wartość AUC	0,817	0,771



5 Podsumowanie

Modele regresji logistycznej oraz wektorów podpierających osiągnęły dobre wyniki w problemie klasyfikacji na naszych danych. Można zauważyć, że w tym przypadku regularyzacja nie wpływała znacząco na jakość predykcyjną modelu. Prawdopodobnie można byłoby ją jeszcze poprawić, optymalizując większą liczbę parametrów.