Sprawozdanie PD2 WdUM

Jakub Niemyjski

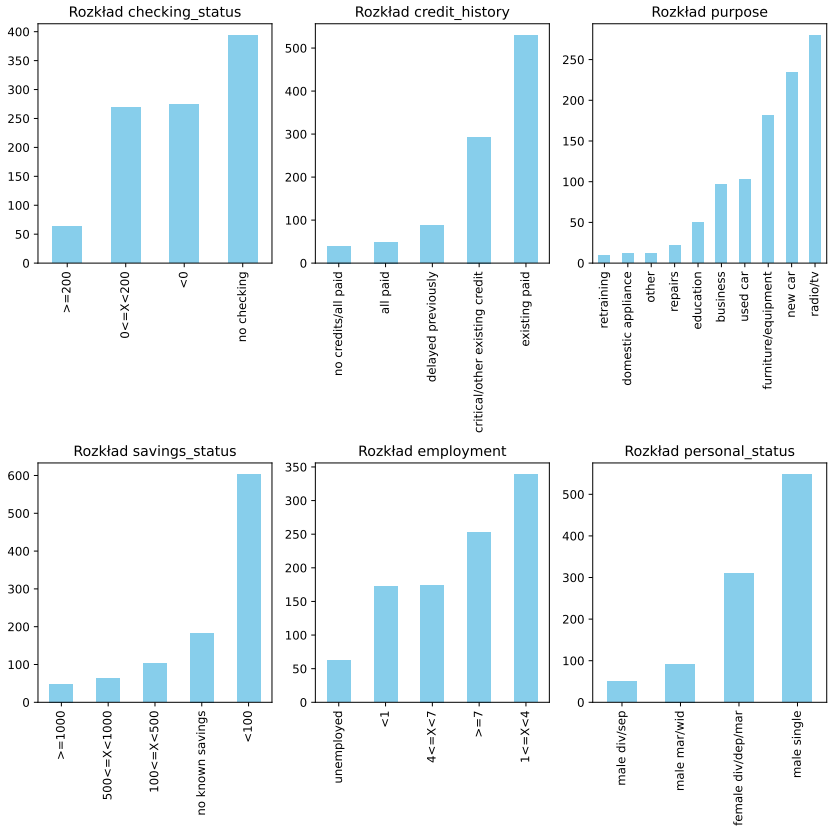
22 listopada 2023

1 Cel

Praca ta ma na celu wykorzystanie modelu regresji logistycznej bez regularyzacji, bądź z regularyzacją l_1- lasso lub l_2 - grzebietową oraz modelu liniowego klasyfikatora SVM do sprawdzenia ich efektywności na rzeczywistych danych.

2 Dane

W tym celu posłużymy się zbiorem danych `credit-g`, który klasyfikuje osoby opisane poprzez zbiór atrybutów jako osoby, którym może być udzielony kredyt lub nie.
Niech `X` oznacz zbiór atrybutów, a `y` będzie oznaczał zmienną celu. Widać, że zbiór `X` posiada 13 kolumn kategorycznych oraz 7 kolumn typu całkowitoliczbowego. Spójrzmy na rozkład kilku atrybutów kategorycznych:



Rysunek 1: Rozkłady sześciu pierwszych zmiennych kategorycznych.

Widać, na przykład, że w naszym zbiorze większość stanowią mężczyźni, którzy nigdy nie mieli partnerki życiowej, a najczęstszym celem kredytów są media takie jak telewizja i radio, bądź nowy samochód lub wyposażenie domu. W dodatku, o zbiorze zmiennych celu można powiedzieć, że 70% badanym osobom przyznano kredyt.

3 Zamiana danych

W celu wykorzystania modelu regresji logistycznej, zmieńmy kolumny kategoryczne na kolumny liczbowe. Niewłaściwym byłoby tu przypisanie kolejnym kategoriom różnych liczb, bo wtedy odnieść możnaby wrażenie, że pewna klasa jest "lepsza" od innej. Aby uniknąć takich sytuacji korzystamy z One Hot Encoding. W rezultacie tworzymy nową ramkę danych `df` złożoną z tej samej liczby wierszy, co posiada `X` i z 59 kolumn.
Do badania jakości modeli, które stworzymy, dzielimy cały zbiór na zbiór danych uczących i zbiór testowy tak, że zbiór testowy stanowi 23% wszystkich obserwacji.

4 Regresja logistyczna

4.1 Brak regularyzacji

Budując model regresji logistycznej bez regularyzacji na zbiorze treningowym, dostajemy taki model, którego krzywa ROC prezentuje się tak, jak pokazano na rysunku 2a, w zależności od tego, czy oceniamy ją na zbiorze treningowym, czy testowym.
Widać, że każdy z estymatorów jest na pewno lepszy niż zupełnie losowy, bo wykres krzywej ROC leży powyżej osi $y = x$. Jednakże wartość AUC dla zbioru treningowego jest dużo lepsza, niż dla zbioru testowego, co może wskazywać na to, że model jest przetrenowany.

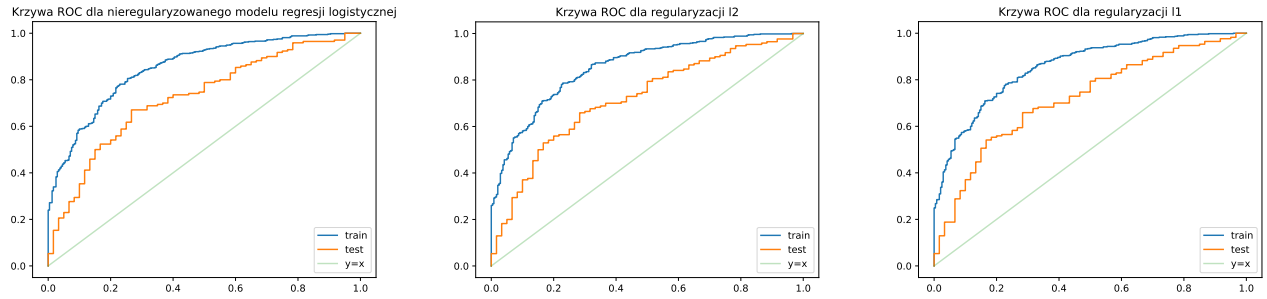
4.2 Regularyzacja Ridge

Aby dokonać najbardziej odpowiedniej regularyzacji l_2, testujemy, dla jakiego parametru C spośród kilku wybranych model zachowuje się najbardziej optymalnie, mierząc efektywność przy wykorzystaniu kroswalidacji. Decyzja o wybraniu konkretnej miary efektywności modelu zależy tu oczywiście od tego, jaki jest cel biznesowy firmy, dla której taką analizę robimy, jednak osobiście uznałem, że najważniejsze jest to, aby nie udzielać zgody na otrzymanie pożyczki osobom, które nie będą w stanie jej spłacić. Stąd odpowiednią miarą wydajności jest precyzja.

4.3 Regularyzacja LASSO

Analogicznie jak w poprzednim podrozdziale, badamy dla jakiego C model będzie miał najwyższą precyzję.

4.4 Zestawienie wyników



(a) Dla modelu nieregularyzowanego (b) Dla modelu z regularyzacją l_2 (c) Dla modelu z regularyzacją l_1

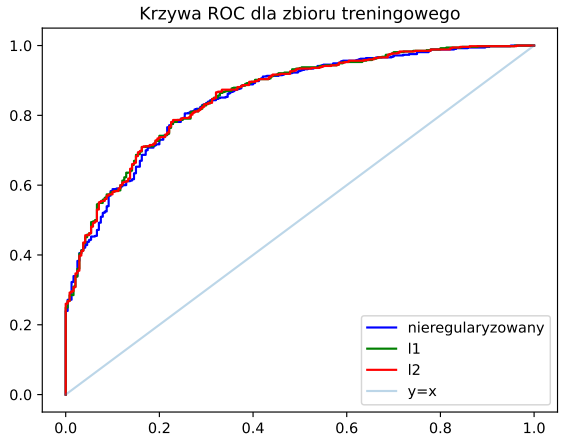
Rysunek 2: Krzywe ROC

Wykresy wyglądają dosyć podobnie. Dla porównania spojrzeć można na przedstawienie tego samego w inny sposób przedstawiony na rysunku 3.

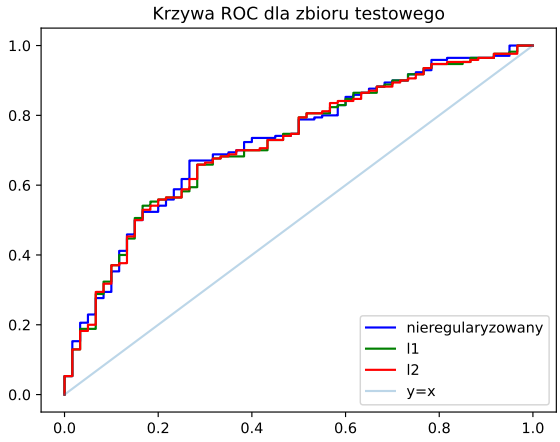
Ponadto, gdy spojrzymy na miary efektywności uzyskanych modeli, czyli na tabelę 4, to faktycznie uznać można, że wszystkie modele są dosyć do siebie zbliżone.

4.5 Redukcja atrybutów

Można przypuszczać, że zmienne, których odpowiadające współczynniki są najmniejsze w regularyzacji l_1, są najmniej ważne w całej estymacji. Ustalono więc pewien próg odcięcia, taki, że jeżeli co do modułu współczynnik był mniejszy od wartości tego progu, to kolumny tej nie braliśmy pod uwagę w budowie nowego modelu z regularyzacją l_1. Analogicznie jak w sekcji 4.2 i 4.3 dobrany został najbardziej odpowiedni parametr C. Model ten nie bierze



(a) Dla zbioru treningowego



(b) Dla zbioru testowego

Rysunek 3: Krzywe ROC

penalty	dokładność	precyzja	czułość	AUC
none	0.800	0.831	0.891	0.745
l1	0.803	0.836	0.887	0.752
l2	0.803	0.836	0.887	0.752

(a) Dla zbioru uczącego

penalty	dokładność	precyzja	czułość	AUC
none	0.704	0.807	0.788	0.627
l1	0.717	0.818	0.794	0.647
l2	0.717	0.818	0.794	0.647

(b) Dla zbioru testowego

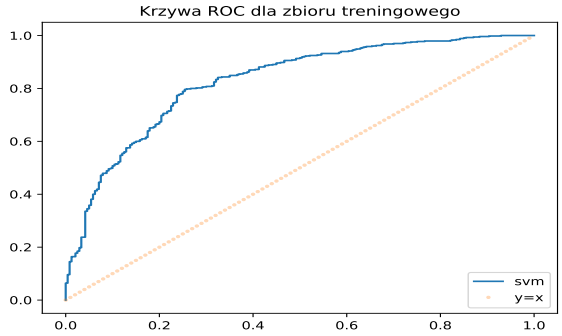
Rysunek 4: Miary efektywości dla modelu regresji logistycznej.

pod uwagę 9 kolumn i do tego ma godziwą precyzję w porównaniu z dotychczas zbudowanymi modelami. Oto jego parametry dla zbioru testowego: dokładność 0.722, precyzja 0.815, czułość 0.806 i AUC 0.645.

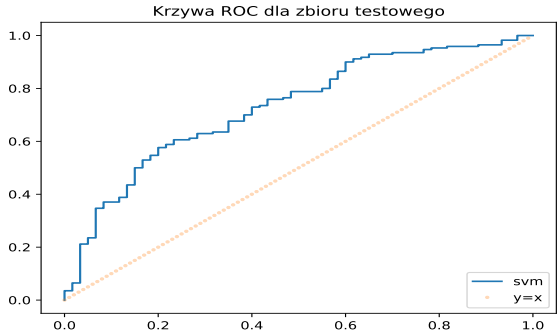
5 SVM

Dla zbioru z usuniętymi 9 kolumnami, o którym była mowa w 4.5, tworzymy model SVM, bez dodatkowej optymalizacji parametru C. Spowodowane zostało to czasem takiej optymalizacji, który był nad wyraz długi. Wyniki przedstawiono w poniższej tabeli i na wykresie 5.

zbiór	dokładność	precyzja	czułość	AUC
uczący	0.790	0.822	0.887	0.731
testowy	0.713	0.817	0.788	0.644



(a) Dla zbioru treningowego



(b) Dla zbioru testowego

Rysunek 5: Krzywe ROC w SVM

6 Podsumowanie

Zapoznaliśmy się z regresją logistyczną oraz modelami SVM. Jak widać, różne metody regularyzacji dawały dosyć zbliżone wartości różnych miar efektywności modeli.
Jeżeli zaś chodzi o model SVM, powiedzieć z pewnością można, że zbudowanie i wytrenowanie go należy (dla większości badanych przeze mnie parametrów C) do czasochłonnych, a akurat przy tych danych dawał on zbliżoną efektywność modelu, co tworzone wcześniej modele regresji logistycznej.