Zastosowanie uczenia maszynowego do predykcji na danych kredytobiorców

Jan Pogłód

1. Wstęp

Celem pracy było zbadanie skuteczności różnych modeli regresji logistycznej na danych kredytobiorców z pewnego banku, oraz wykorzystanie tej wiedzy do określenia czy są oni zdolni do kredytu. W eksperymencie badane były modele regresji logistycznej z regularyzacją z normami L1, L2 oraz bez reguraryzacji. Następnie właściwości jakie niosły ze sobą te modele posłużyły do eliminacji niektórych zmiennych i przygotowania modelu wektorów podpierających. Poniżej krok po kroku przedstawiłem kroki osiągnięcia najlepszych modeli.

2. Przygotowania najlepszych modeli regresji logistycznej

Za pomocą 5-krotnej kroswalidacji krzyżowej oraz funkcji grid_search przeprzeprowadziłem badanie mające na celu znalezienie najlepszych współczynników do powyższych modeli regresji logistycznych. Wyniki podzieliłem poniżej względem miar dokładności, czułości i precyzji, miara AUC. Następnie w kolejnym punkcie osobno przeanalizowałem skuteczność krzywej ROC dla każdego z tych modeli.

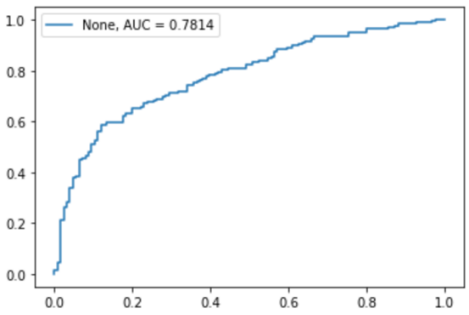
2.1 Wyniki na zbiorze treningowym

	Regularyzacja L1	Regularyzacja L2	Bez regularyzacji
Dokładność	0.8167	0.7983	0.8050
Precyzja	0.8465	0.8337	0.8411
Czułość	0.9061	0.8944	0.8944
Miara AUC	0.8443	0.8566	0.8466

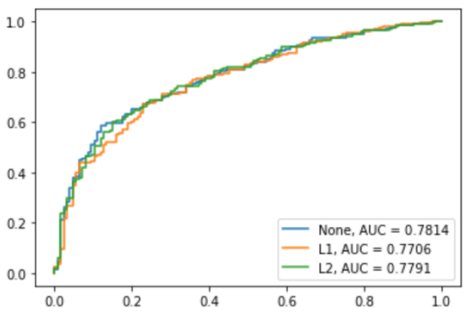
2.2 Wyniki na zbiorze testowym

	Regularyzacja L1	Regularyzacja L2	Bez regularyzacji
Dokładność	0.7250	0.7325	0.7375
Precyzja	0.7645	0.7651	0.77
Czułość	0.8650	0.8796	0.8796
Miara AUC	0.7791	0.7714	0.7814

Analizując powyższe tabele widzimy, że na zbiorze treningowym najlepiej poradził sobie model z regularyzacją L1, osiągając najwyższe wyniki dokładności, precyzji i czułości. Lepiej od niego wypadł jedynie model z regularyzacją L2 względem miary AUC. Jeżeli spojrzymy na wyniki na zbiorze testowym bezwzględnie króluje wśród nich model bez regularyzacji osiągając najlepsze wyniki w każdej mierze dokładności, precyzji, czułości i miary AUC. Także dla modelu bez regularyzacji wykres krzywej ROC okazał się być najskuteczniejszy, co możemy zobaczyć poniżej:



Rys. 1. Krzywa ROC dla modelu regresji logistycznej bez regularyzacji – najlepszy model



Rys. 2. Krzywa ROC dla wszystkich modeli regresji logistycznej wraz z miarami AUC

Podsumowując powyższe wyniki, ze wszystkich trzech modeli najskuteczniejszymi względem jakości predykcji został model regresji logistycznej bez regularyzacji. Osiągnął on miarę dokładności na poziomie *0.7375* oraz miarę AUC na poziomie *0.7814*. Na zbiorze treningowym najlepszym okazał się model z regularyzacją L1. Oba z tych modeli wykorzystam teraz do eliminacji zmiennych, które mogą mieć mniej istotny wpływ na model.

3. Zmienne o mniejszym wpływie

Ramka danych zawierające dane klientów banku miała 21 atrybutów: 7 atrybutów numerycznych i 14 atrybutów kategorycznych. Za pomocą funkcji z pakietu pandas *get_dummies* zamieniłem zmienne kategoryczne na binarne dla każdej kategorii (odrzucając za każdym razem jedną kategorię, ponieważ jest ona liniowo zależna od pozostałych) otrzymując w ten sposób ramkę danych z 48 atrybutami. Na początek przeanalizowałem, które ze zmiennych kategorycznych są możliwe do usunięcia.

3.1 Usunięcie części zmiennych kategorycznych

Na początku, korzystając ze współczynników obliczonych za pomocą najlepszego modelu regresji logistycznej z poprzedniej części - bez regularyzacji oraz z modelu z regularyzacją L1 odkryłem, że niektóre zmienne kategoryczne mają większy wpływ na rezultat otrzymania kredytu od innych. W tych zmiennych, o największych wpływie są na przykład:

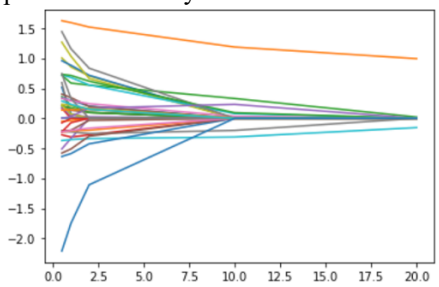
Zmienne wykryte przez model bez regularyzacji:

- [Cel] Nowy samochód: **-0.6339**
- [Cel] Używany samochód: **1.2374**
- [Zatrudnienie] 4 – 7 lat: **0.5531**
- [Rachunki bankowe] Brak: **1.5614**
- [Oszczędności] Więcej niż 1000: **0.8082**
- [Status] Samotny mężczyzna: **0.6182**

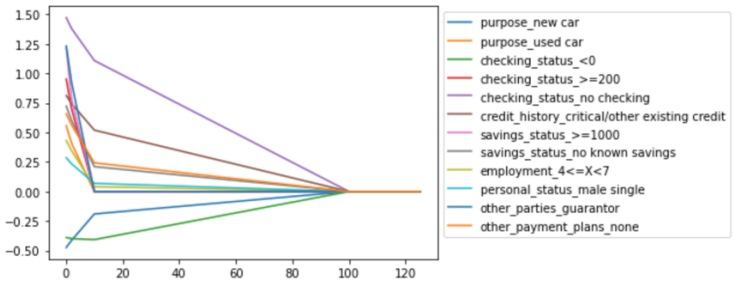
Zmienne wykryte przez model z regularyzacją L1:

- [Cel] Przebranżowienie: **3.0739**
- [Zatrudnienie] Za granicą: **-1.8336**
- [Zatrudnienie] bezrobotny: **-0.6024**
- [Zamieszkanie] Wyanjmowany dom: **-0.6131**
- [Oszczędności] Więcej niż 1000: **2.0344**
- [Historia kredytów] Poprzednio opóźniony: **0.8983**

Mimo, że oba modele zakwalifikowały te same zmienne, które wpływają pozytywnie na dostanie kredytu (wartości dodatnie) i te które wpływają na to negatywnie (wartości ujemne) i wartości te się dużo nie różniły, to część zmiennych okazała się dać inne wyniki. Dzięki modelowi bez regularyzacji okazało się, że jest duża różnica, czy kredyt bierzemy na samochód nowy czy używany, samotny mężczyzna ma większe na niego szanse i brak innych rachunków bankowych ma największy wpływ na pozytywną ocenę kredytodawcy. Z modelu z regularyzacją L1 okazało się za to, że cel kredytu „Przebranżowienie" jest najmilej widziany przez banki, zaś bezrobotni, pracujący za granicą lub zamieszkali w wynajmowanych domach mają znacznie mniejszą szansę na dostanie pieniędzy. W obu modelach wielki wpływ ma także posiadanie oszczędności większych niż 1000. Przeanalizowałem dokładnie zmienne bardziej wpływowe i mniej i dostałem poniższe rezultaty:



Rys 3. Wszystkie zmienne przy różnych parametrach C, przed usunięciem mało istotnych



Rys 4. Najbardziej wpływowe zmienne po usunięciu tych, które w regularyzacji zerują się na poziomie C = 2 w modelu regresji

Na wykresie 3 przedstawiłem wszystkie zmienne kategoryczne oraz skuteczność z jaką model z regularyzacją L1 zerował współczynniki przy nich dla kolejnych parametrów C. Widzimy, że dla niektórych zmiennych współczynniki przy nich wyzerowały się już na poziomie parametrów *C=1, 2, 3* oraz *10*. Jako parametr do dyskwalifakcji zmiennych użyłem paramatru *C=2*. Na wykresie 4, widzimy rezultat usunięcia tych zmiennych oraz nazwy, tych najbardziej znaczących, które zostały. Następnie przebadałem otrzymane zmienne, aby sprawdzić czy można jeszcze zrezygnować z jakiś zmiennych na przykład przez podobieństwo do innych. Przeprowadziłem ponowne badanie i okazało się, że zmienne takie jak *[Cel]: Przekwalifikowanie, [Historia kredytowa]: Brak/ Opłacone, [Historia kredytowa]: Opóźnienie* mają współczynniki na poziomie *0,00-0,05*, a więc stosunkowo bardzo niskie do pozostałych. Przeprowadziłem test studenta dla tych zmiennych wraz z odpowiednimi jak na dole:

Kategoria	Zmienna 1	Zmienna 2	P-value
Historia/ Oszczędności	Brak kredytów	Więcej niż 1000	0.3833
Historia/ Status	Opóźniony	Żonaty	0.756
Cel/ Cel	Przekfalifikowanie	Inny	0.5107

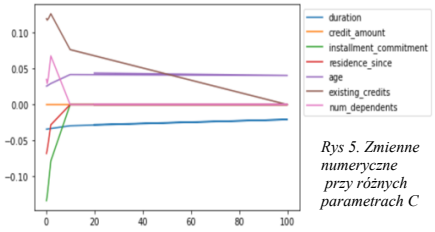
Hipoteza zerowa: Nie ma istotnej różnicy między dwoma zmiennymi (grupami)

P-value < 0.05: Są podstawy do odrzucenia hipotezy zerowej

W przypadku rozważanych zmiennych oraz odpowiadającym im zmiennym do porównania dostaliśmy P-value na poziomie znacząco większym niż alfa=0.05, co tym bardziej potwierdza przypuszczenia, że powyższe zmienne mogą być mniej istotne dla budowy modelu. Na tej podstawie podjąłem decyzje o usunięciu również tych zmiennych. Przeanalizowałem zależność wyniku tstudenta również dla innych zmiennych, ale wyniki te okazały dawać podstawy do odrzucenia hipotezy zerowej. Uznałem więc, że pozostałe zmienne są istotnie różne statystycznie.

3.2 Usunięcie części zmiennych numerycznych

W celu rozważenia czy któreś zmienne z tych o typie numerycznym mogą zostać poddane usunięciu przeprowadziłem ponowne rozważania jak w przypadku powyższej analizy. Na wykresie 5, za pomocą modelu regresji logistycznej z regularyzacją L1 widzimy, że są trzy współczynniki dla zmiennych *[Ilość osób na utrzymaniu], [Zobowiązanie do zapłaty] i [Zamieszkanie od],* które eksperyment wyzerował już na poziomie parametru C=10, tym samym wyróżniając je na tle pozostałych. Przeanalizowałem te zmienne dokonując korelacji Pearsona wraz z pozostałymi i okazało się, że mają one bardzo niską korelacje Pearsona, a zatem niosą wiele cech induwidualnych dla modelu. Jedynie zmienne *[Okres kredytu]* i *[Wysokość kredytu*] oraz *[Wysokość kredytu]* i *[Zobowiązanie do zapłaty]* są silniej skolerowane, na poziomie odpowiednio 0.625 i -0.271, ale wykres współczynników daje, że są one znaczące dla modelu, mimo ich niskiej wartości względnej współczynników.
Na podstawie powyższego nie podjąłem decyzji o usunięciu jakiejkolwiek ze zmiennych numerycznych.



Rys 5. Zmienne numeryczne przy różnych parametrach C

Ilość osób na utrzymaniu Korelacja względem:	
Okres kredytu	-0.0238
Wysokość kredytu	0.0171
Zobowiązanie do zapłaty	-0.0712
Zamieszkanie od	0.0426
Wiek	0.1182

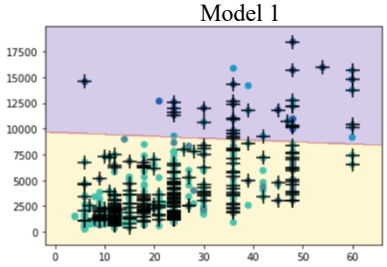
Wysokość kredytu Korelacja względem:	
Okres kredytu	**0.625**
Zobowiązanie do zapłaty	**-0.2713**
Zamieszkanie od	0.0289
Inne kredyty	0.0208
Osoby na utrzymaniu	0.0171

Tabele przedstawiające siłę skolerowania wybranych zmiennych

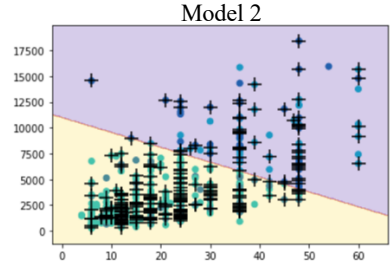
4. Model wektorów podpierających SVM

Po badaniu które zmienne mogą zostać usunięte stworzyłem nową ramkę danych, na podstawie konkatenacji dwóch powstałych ramek – z atrybutami numerycznymi i kategorycznymi, która ostatecznie zawierała już 23 wyselekcjonowane zmienne, a nie 48 jak początkowa. Tą ramkę podzieliłem na zbiory testowe i treningowe, a następnie wytrenowałem na nich model wektorów podpierających SVC. Następnie korzystając z funkcji grid search znalazłem najlepszy z hiperparametrów C, który wyniósł 5. Wytrenowany w ten sposób model nazwałem poniżej w ramce podsumowującej modelem 1. Następnie dokonałem innej selekcji zmiennych, tym razem w celach testowych zakładając, że odrzucać będę te zmienne kategoryczne, których względna wartość współczynników jest mniejsza od 0.5 dla regresji logistycznej bez regularyzacji, która okazała się najlepsza w części pierwszej. Model ten nazwałem model 2 i porównawcze wyniki zaprezentowałem poniżej.

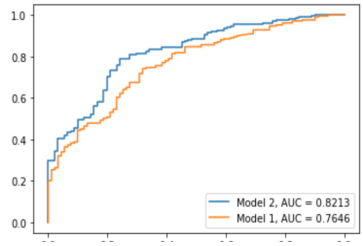
Model 1



Model 2



Krzywe ROC



Model 1	Dokładność	AUC	Precyzja	Czułość
Testowy	**73,67%**	**76,46%**	**77,78%**	**87,08%**
Treningowy	**75,14%**	**78,26%**	**78,98%**	**87,98%**

Model 2	Dokładność	AUC	Precyzja	Czułość
Testowy	**77,33%**	**82,13%**	**81,33%**	**87,56%**
Treningowy	**79,11%**	**83,09%**	**82,83%**	**88,17%**

5. Podsumowanie

Ostateczne modele wektorów podpierających osiągnięte różnymi technikami okazały się mieć zbliżone, lecz różne wyniki. Lepszy w tym wariancie okazał się być model, w którym zmienne usunięto na podstawie zależności współczynników najlepszego modelu regresji logistycznej z poprzedniej części – bez regularyzacji. Mimo tego warto jest dokonywać różnych analiz modelów regresji logistycznych do dalszych badań, aby zobaczyć, na przykładzie modelu z regularyzacją L1, które współczynniki w jakim tempie dążą do zera. Gwatantuje to najlepszą selekcję zmiennych do końcowych wyników.