Wstęp do uczenia maszynowego
Praca domowa 2

Franciszek Saliński

19 listopada 2024

1 Eksploracja i przygotowanie danych

Analizując dane sprawdziłem, czy występują brakujące wartości, których nie znalazłem. Następnie sprawdziłem typy danych oraz przyjrzałem się rozkładom i macierzy korelacji.
Przetworzyłem dane, tak aby nadawały się do modelowania. Dla kolumny 'employment' zastosowałem label encoding przypisując kolejnym przedziałom rosnąco wartości 0, 1, 2, 3 i 4. Dla pozostałych kolumn kategorycznych zastosowałem One-hot encoding wyrzucając jedną z kolumn. Kolumny numeryczne ustandaryzowałem do rozkładu ze średnią 0 i odchyleniem standardowym 1. Zmienną celu zmapowałem tak, żeby wartość 'bad' była 1, a 'good' 0, ponieważ bardziej interesuje nas model który skutecznie wykrywał będzie tych klientów, którym nie warto dać kredytu. Zmienna celu nie jest idealnie zbalansowana, dlatego dla wszystkich modeli używałem parametru $class\_weigth = "balanced"$.
Podzieliłem dane na zbiór treningowy i testowy w stosunku 8 : 2.

2 Część 1

Za pomocą 5-krotnej kroswalidacji na zbiorze treningowym szukałem modeli, które osiągną najlepszą metrykę AUC. W tym celu optymalizowałem hiperparametr C dla modeli z regularyzacją L1, L2 oraz ElasticNet, a także $l1\_ratio$ dla modelu ElasticNet, sprawdzając wartości:

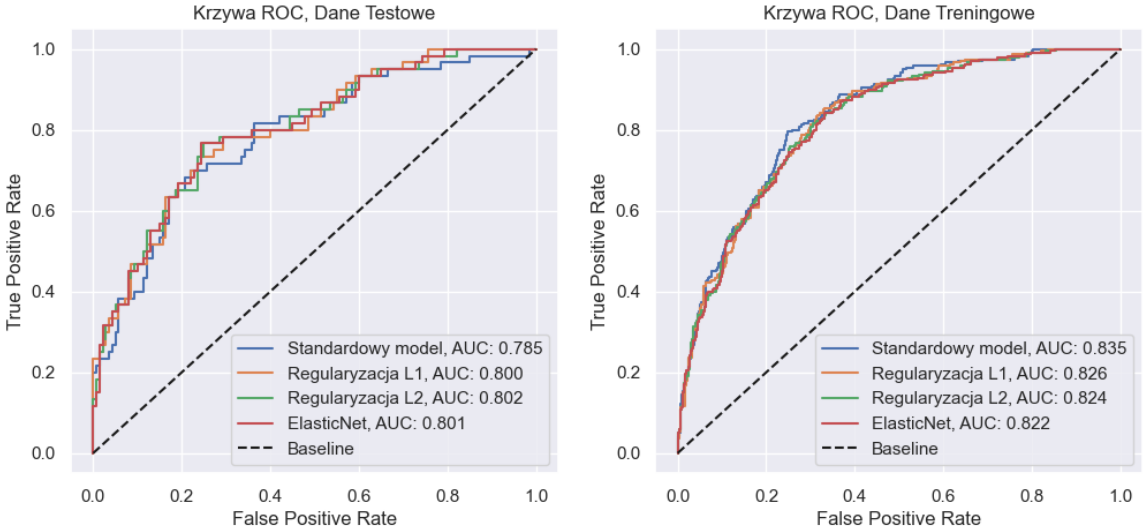
- $C = [0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 10, 20]$
- $l1\_ratio = [0.1, 0.25, 0.5, 0.75, 0.9]$

Dla modelu ElasticNet zastosowałem też większą niż standardowe 100 maksymalną liczbę iteracji, ponieważ dla 100 iteracji miał problem ze zbieżnością, co prawdopodobnie wynika z solvera 'saga', który jest jedynym dostępnym dla tego modelu. Najlepsze uzyskane modele:

- Model z regularycją L1 (dalej 'L1') z parametrem $C = 0.5$
- Model z regularycją L2 (dalej 'L2') z parametrem $C = 0.1$
- Model z regularycją ElasticNet (dalej 'ElasticNet') z parametrami $C = 0.1$ i $l1\_ratio = 0.1$

Następnie modele ze znalezionymi optymalnymi parametrami wytrenowałem na całym zbiorze treningowym, tak samo jak model bez regularyzacji (dalej 'No penalty'). Te 4 modele walidowałem na zbiorze testowym i uzyskałem następujące metryki oraz krzywe ROC:

Model	Data	Accuracy	Precision	Recall	F1	AUC
No penalty	Test	0.735	0.545	0.700	0.613	0.785
No penalty	Train	0.760	0.574	0.775	0.660	0.835
L1	Test	0.730	0.537	0.733	0.620	0.800
L1	Train	0.741	0.550	0.758	0.637	0.826
L2	Test	0.750	0.562	0.750	0.643	0.802
L2	Train	0.746	0.557	0.758	0.642	0.824
ElasticNet	Test	0.750	0.561	0.767	0.648	0.801
ElasticNet	Train	0.738	0.545	0.754	0.633	0.822



Najlepiej moim zdaniem radzi sobie model ElasticNet, uzyskał najlepszą czułość (Recall) na zbiorze testowym oraz ma najmniej rozbieżne metryki między zbiorem treningowym, a testowym. Czułość na poziomie 0.767 z jednoczesną precyzją 0.561 oznacza, że model wykrywa około 77% wszystkich tych klientów, którym nie powinniśmy udzielić kredytu i w predykcji tej klasy ma skuteczność 56%.

Przyjrzałem się także wielkościom współczynników każdego z modeli, tabela ze współczynnikami dla każdej kolumny znajduje sie na ostatniej stronie raportu. Możemy dojść do wniosku, że najistotniejszymi zmiennymi w predykcji były:

- 'checking_status_no checking'
- 'credit_history_critical/other existing credit '
- 'purpose_education'
- 'purpose_used car
- 'foreign_worker_yes'

przy czym, z powyższych zmienne 'foreign_worker_yes' oraz 'purpose_education' sprzyjały predykcji klasy 1, pozostałe 0. Jeśli chodzi o zmienne mało istotne, to zająłem się nimi w 2 części zadania.

3 Część 2

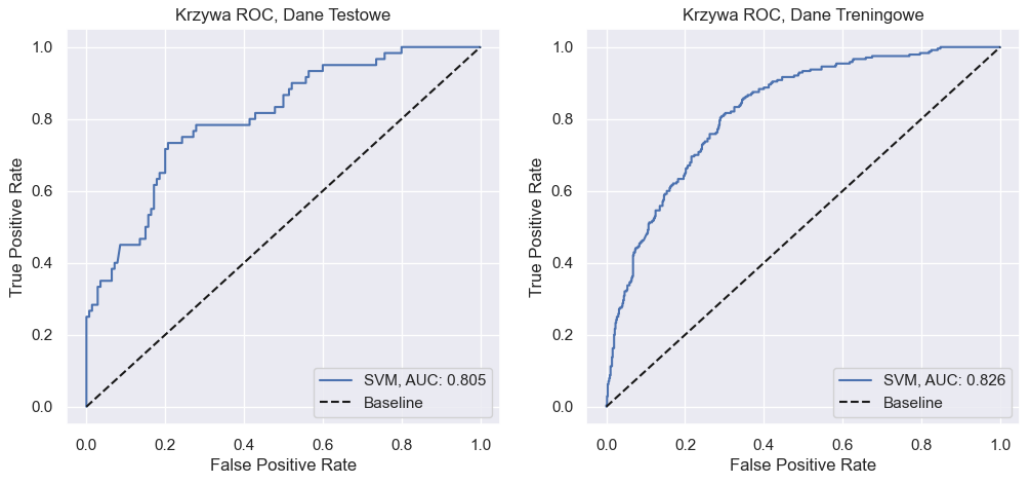
Aby zredukować wymiar danych postanowiłem się posłużyć współczynnikami modelu z regularyzacją L1, który jak wiemy może posłużyć jako narzędzie wyboru zmiennych. Sprawdziłem które współczynniki w modelu L1 były równe 0, zdecydowałem, że można te zmienne uznać za mało istotne i pozbyłem się w ten sposób 14 kolumn.

Następnie już ze zredukowanymi danymi postepowałem podobnie jak w części 1. W modelu SVC (dalej SVM Linear) zdecydowałem się na jądro liniowe, a w 5-krotnej kroswalidacji szukałem możliwie najlepszego parametru C sprawdzając wartości:

- $C = [0.01, 0.02, 0.05, 0.1, 0.2, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 5, 10, 20]$

Najlepszy wynik AUC = 0.785 uzyskałem dla parametru $C = 0.2$. Wytrenowałem model z takim parametrem na całym zbiorze treningowym i uzyskałem następujące wyniki:

Model	Data	Accuracy	Precision	Recall	F1	AUC
SVM Linear	Test	0.735	0.542	0.750	0.629	0.805
SVM Linear	Train	0.732	0.538	0.767	0.632	0.826



Jakość predykcyjna modelu jest bardzo podobna do regresji logistycznej z regularyzacją ElasticNet. Dostaliśmy niewiele mniejsze wartości metryk recall i precision, które są moim zdaniem najbardziej istotne.

Feature	No penalty	L1	L2	ElasticNet
duration	0.340	0.316	0.316	0.315
credit_amount	0.388	0.339	0.275	0.271
employment	-0.181	-0.148	-0.158	-0.152
installment_commitment	0.419	0.362	0.327	0.320
residence_since	0.026	0.013	0.024	0.020
age	-0.168	-0.147	-0.164	-0.154
existing_credits	0.211	0.133	0.101	0.096
num_dependents	0.126	0.093	0.095	0.087
checking_status_<0	0.141	0.213	0.308	0.298
checking_status_>=200	-0.209	0.000	-0.053	-0.005
checking_status_no checking	-1.462	-1.289	-0.993	-0.996
credit_history_critical/other existing credit	-1.726	-0.896	-0.628	-0.598
credit_history_delayed previously	-1.163	-0.154	-0.098	-0.031
credit_history_existing paid	-0.925	-0.189	-0.088	-0.053
credit_history_no credits/all paid	-0.525	0.000	0.163	0.131
purpose_domestic appliance	-0.114	0.000	-0.023	0.000
purpose_education	1.054	0.728	0.378	0.349
purpose_furniture/equipment	-0.068	0.000	-0.051	-0.015
purpose_new car	0.677	0.583	0.422	0.411
purpose_other	-0.639	0.000	-0.035	0.000
purpose_radio/tv	-0.246	-0.187	-0.230	-0.214
purpose_repairs	0.708	0.074	0.154	0.092
purpose_retraining	-5.447	0.000	-0.148	-0.061
purpose_used car	-1.104	-0.749	-0.541	-0.505
savings_status_500<=X<1000	0.299	0.000	-0.009	0.000
savings_status_<100	0.581	0.508	0.392	0.390
savings_status_>=1000	-0.447	0.000	-0.162	-0.104
savings_status_no known savings	-0.375	-0.349	-0.359	-0.338
personal_status_male div/sep	0.234	0.053	0.162	0.114
personal_status_male mar/wid	-0.393	-0.250	-0.214	-0.178
personal_status_male single	-0.585	-0.438	-0.368	-0.348
other_parties_guarantor	-0.893	-0.330	-0.228	-0.192
other_parties_none	-0.193	0.000	0.075	0.045
property_magnitude_life insurance	-0.007	0.000	-0.018	0.000
property_magnitude_no known property	0.523	0.329	0.257	0.242
property_magnitude_real estate	-0.305	-0.251	-0.266	-0.249
other_payment_plans_none	-0.481	-0.418	-0.353	-0.343
other_payment_plans_stores	0.097	0.000	0.104	0.056
housing_own	0.273	0.000	-0.121	-0.123
housing_rent	0.670	0.317	0.184	0.155
job_skilled	-0.336	0.000	-0.092	-0.053
job_unemp/unskilled non res	-1.187	-0.007	-0.159	-0.077
job_unskilled resident	-0.320	0.000	-0.035	0.000
own_telephone_yes	-0.552	-0.340	-0.306	-0.277
foreign_worker_yes	1.289	0.933	0.354	0.314