1 Przygotowanie zbioru danych

Podczas przygotowania danych potwierdziłam brak brakujących wartości, ale należało dokonać transformacji zmiennych kategorycznych. Nierównomierny rozkład zmiennej celu uwzględniłam poprzez stratyfikowaną kroswalidację i użycie parametru stratify przy podziale danych. Skośność zmiennych numerycznych zniwelowałam transformacją logarytmiczną.

2 Część 1.

2.1 Model Regresji Logistycznej bez regularyzacji

Model regresji logistycznej uwzględniał dodatkowy parametr liczby iteracji ze względu na pojawiające się ostrzeżenie o braku zbieżności algorytmu lbfgs.
Przeanalizowałam tylko solver 'lbfgs'. Pozostałe z samej definicji zawartej w dokumentacji preentowały się jako niedostosowane do tego zadania.

Model osiągnął lepsze wyniki na danych treningowych niż testowych, co wskazuje na przetrenowanie spowodowane bra-

wartości metryk	zbiór testowy	zbiór treningowy
dokładność	0.727	0.796
czułość	0.862	0.904
precyzja	0.774	0.822
miara AUC	0.781	0.833

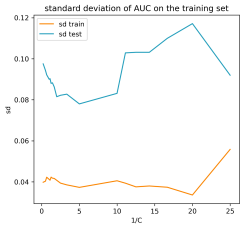
kiem regularizacji. Czułość jest wysoka (model dobrze identyfikuje przypadki pozytywne), ale niższa dokładność, precyzja i AUC na danych testowych wskazują na błędy predykcyjne. Wprowadzenie regularizacji mogłoby poprawić ogólną jakość modelu.

2.2 Model Regresji Logistycznej z Regularyzacją

Analizę wpływu regularyzacji na model regresji logistycznej rozpoczęłam od przygotowania funkcji, która przeprowadzała 10-krotną stratyfikowaną kroswalidację, wyznaczając w ten sposób miarę AUC na zbiorze treningowym i testowym. Funckja zwraca wektor tych miar, co pozwala na wyznaczenie wartości średniej oraz odchyelnia standardowego, które są kluczowe w dalszej analizie.

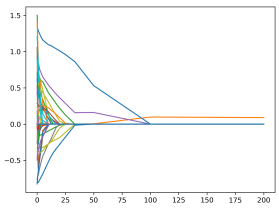
2.2.1 Regularyzacja L1

Przygotowanie jak najlepszego modelu regresji liniowej z regularyzacją L1 rozpoczęłam od analizy miary wartości AUC wykorzystując przygotowaną funkcję.

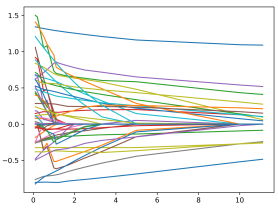


Rysunek 1:

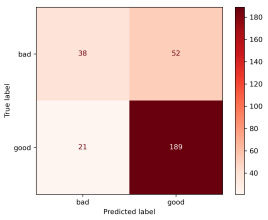
Rysunek 2:



Rysunek 3:



Rysunek 4:



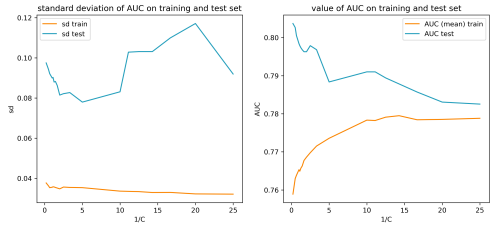
W przedziale $\frac{1}{C}$ od 1 do 5 (rys. 1) model osiąga niskie odchylenie standardowe i maksymalne AUC na zbiorze testowym, co wskazuje na stabilność i najlepszą generalizację. Analiza współczynników (rys. 2,3) potwierdza, że regularyzacja L1 skutecznie selekcjonuje istotne cechy, stabilizując wartości w tym przedziale i równoważąc przeuczenie z niedouczeniem. Grid Search potwierdziło optymalność parametru C = 0.3 , zgodnie z wcześniejszymi wynikami.

wartości metryk	zbiór testowy	zbiór treningowy
dokładność	0.757	0.767
czułość	0.9	0.918
precyzja	0.784	0.785
miara AUC	0.804	0.813

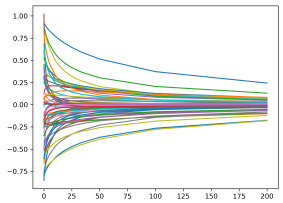
Model z regularyzacją L1 osiągnął wysoką skuteczność bez przetrenowania, z czułością 0.9 i AUC 0.804 na zbiorze testowym. Dobrze identyfikuje klasę "good" (189 poprawnych predykcji), ale ma trudności z klasą "bad" (38 poprawnych, 52 błędnych). Wyniki wskazują na dobrą separację klas i stabilność modelu.

2.2.2 Regularyzacja L2

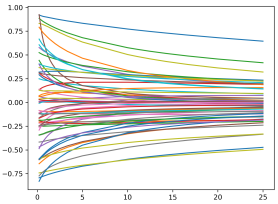
W przypadku regularyzacji L2 postępowałam analogicznie jak wybierając możliwie najlepszy model z regularyzacją L1.



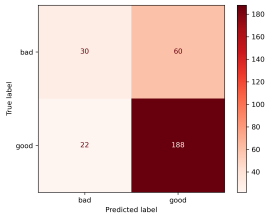
Rysunek 5:



Rysunek 6:



Rysunek 7:



Rysunek 8:

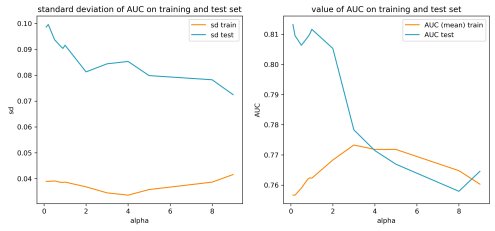
Zakres $\frac{1}{C}$ (rys. 5) od 0 do 2 jest optymalny dla regularyzacji L2, ponieważ zapewnia stabilność modelu, maksymalizuje AUC testowe i minimalizuje różnice między zbiorami treningowym a testowym, unikając przetrenowania. Regularyzacja w tym zakresie ogranicza wariancję, zachowując zdolność generalizacji i skutecznie wykorzystując cechy. Grid Search potwierdziło, że optymalna wartość C wynosi 0.07, zgodnie z wcześniejszymi analizami.

wartości metryk	zbiór testowy	zbiór treningowy
dokładność	0.727	0.764
czułość	0.895	0.929
precyzja	0.758	0.778
miara AUC	0.799	0.821

Model L2 osiągnął równowagę między zbiorami, z dokładnością 0.727 i AUC 0.799 na zbiorze testowym. Czułość 0.895 wskazuje na skuteczne wykrywanie pozytywnych przypadków, a model dobrze generalizuje na nowe dane. Model z regularyzacją L2 skutecznie identyfikuje pozytywne klasy ("good"), co potwierdza 188 poprawnych predykcji. Jednak 60 przypadków klasy "bad" zostało błędnie zaklasyfikowanych jako "good", co wskazuje na pewne problemy z dokładnością w wykrywaniu negatywnych przykładów.

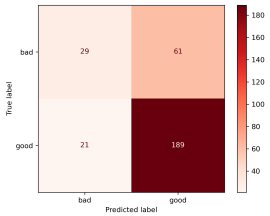
2.3 Model Elastic Net

Ostatnim analizowanym przez mnie modelem był Elastic Net, czyli model regularyzacji, stosowany w regresji, łączyc dwie metody regularyzacji: L1 i L2.
Szukając jak najlepszego modelu skupiłam się na analizie dwóch hiperparametrów: alpha oraz l1 ratio.



Rysunek 9:

Rysunek 10:



Rysunek 11:

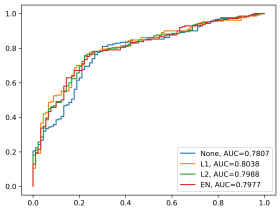
Przedział alpha (rys. 9) od 0 do 2 jest optymalny dla modelu Elastic Net, zapewniając najwyższe AUC na zbiorze testowym i niewielkie różnice między zbiorami. Wskazuje to na dobrą generalizację i równowagę między dopasowaniem a uproszczeniem modelu.
Przedział l1 ratio (rys. 10) od 0.2 do 0.4 jest optymalny, zapewniając najwyższe AUC na zbiorze testowym oraz stabilne odchylenie standardowe na treningowym. Poza tym zakresem model staje się zbyt uproszczony, tracąc zdolność predykcyjną. Ten zakres równoważy stabilność, elastyczność i generalizację.

W celu podsumowania wykonanej analizy ponownie skorzystałam z funkcji Grid Search, aby wybrać najlepszą parę parametrów alpha, l1 ratio. Tym razem dobrałam inne miary, bardziej kompatybilne z modelem Elastic Ney. Otrzymałam, że optymalną parą hiperparametrów są alpha = 1 (C = 0.1) oraz l1 ratio = 0.3 co zgadza się z wybranymi optymalnymi przedziałami dla tych hiperparametrów.

wartości metryk	zbiór testowy	zbiór treningowy
dokładność	0.753	0.764
czułość	0.924	0.931
precyzja	0.77	0.777
miara AUC	0.798	0.813

Model Elastic Net dobrze dopasowuje dane, osiągając wysoką czułość i solidne AUC. Precyzja 0.767 wskazuje na satysfakcjonującą jakość z niewielką liczbą fałszywych alarmów. Minimalne różnice między zbiorami sugerują dobrą generalizację.

2.4 Krzywe ROC



Model z regularyzacją L1 osiąga najwyższe AUC, przewyższając L2 i Elastic Net. Brak regularyzacji wypada najgorzej, co potwierdza korzyści regularyzacji. L1 jest najskuteczniejszym wyborem, łącząc wysoką skuteczność z ograniczeniem przeuczenia.

2.5 Podsumowanie

Model z regularyzacją L1 osiągnął najwyższe AUC, skutecznie eliminując mniej istotne cechy. L2 osiągnął nieco niższe wyniki, a Elastic Net nie przyniósł dodatkowych korzyści. Model bez regularyzacji wypadł najgorzej z powodu przetrenowania.

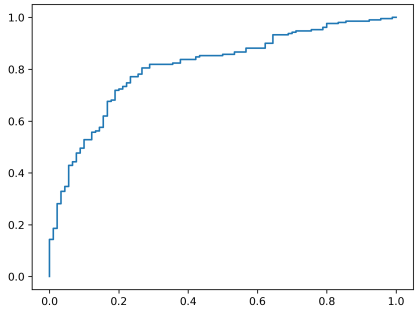
Na podstawie wartości współczynników dla modelu L1 możemy wprowadzić selekcję zmiennych. Możemy ograniczyć ich liczbę z 61 do 24.

3 Część 2.

Analogicznie do rozumowania w części 1. zbadałam i przygotowałam model wektorów podpierających. Zaczęłam od przeprowadzenia stratyfikowanej kroswalidacji, aby wybrać optymalne przedziały dla hiperparametrów. Potwierdziałam następnie moją analizę funkcją Grid Search, która wskazała jako optymalne C: 0.4 oraz kernel: linear.

wartości metryk	zbiór testowy	zbiór treningowy
dokładność	0.733	0.779
czułość	0.881	0.914
precyzja	0.771	0.799
miara AUC	0.811	0.815

Model wektorów nośnych (SVM) osiągnął bardzo dobre wyniki, przewyższając inne analizowane modele. Wartość AUC wynosząca 0.811 na zbiorze testowym i 0.815 na treningowym wskazuje na znakomitą zdolność do rozróżniania klas. Wysoka zgodność między zbiorami sugeruje brak przetrenowania, a precyzja i czułość świadczą o skuteczności w klasyfikacji zarówno pozytywnych, jak i negatywnych przypadków. Model z regularyzacją C=0.4 i liniowym jądrem okazał się wyjątkowo dobrze dopasowany do analizowanego zbioru danych.



Krzywa ROC dla modelu SVM pokazuje bardzo dobrą jakość predykcyjną, z dużą powierzchnią pod krzywą. Wskazuje to na wysoką zdolność modelu do rozróżniania klas i skuteczność w klasyfikacji zarówno pozytywnych, jak i negatywnych przypadków. Model charakteryzuje się świetną równowagą między czułością a specyficznością.

4 Tabela z wartościami współczynników

Nazwa zmiennej	bez regularyzacji	L1	L2	Elastic Net
duration	-0.802	-0.604	-0.476	-1
credit_amount	-0.287	0	-0.234	0
installment_commitment	-0.209	-0.093	-0.182	0
residence_since	-0.024	0	-0.021	0
age	0.123	0.521	0.088	0
existing_credits	-0.218	0	-0.058	0
num_dependents	-0.113	0	-0.031	0
checking_status_0<=X<200	-0.147	0	-0.224	0
checking_status_<0	-0.464	-0.449	-0.569	-1
checking_status_>=200	0.877	0.337	0.173	0
checking_status_no checking	1.17	1.016	0.619	1
credit_history_all paid	-0.705	-0.255	-0.322	0
credit_history_critical/other existing credit	1.398	0.709	0.513	1
credit_history_delayed previously	0.705	0	0.038	0
credit_history_existing paid	0.4	0.124	0.019	0
credit_history_no credits/all paid	-0.362	-0.158	-0.247	0
purpose_domestic appliance	1.159	0	0.067	0
purpose_new car	-0.907	-0.535	-0.416	0
purpose_used car	0.755	0.379	0.322	0
purpose_business	-0.061	0	-0.037	0
purpose_education	-0.469	0	-0.095	0
purpose_furniture/equipment	0.078	0	0.076	0
purpose_other	1.03	0	0.089	0
purpose_radio/tv	-0.181	0	0.061	0
purpose_repairs	-0.902	0	-0.121	0
purpose_retraining	0.935	0	0.053	0
savings_status_100<=X<500	0.685	0	0.129	0
savings_status_500<=X<1000	-0.241	0	-0.057	0
savings_status_<100	-0.28	-0.433	-0.356	0
savings_status_>=1000	0.726	0	0.095	0
savings_status_no known savings	0.548	0.034	0.190	0
employment_1<=X<4	0.211	0	-0.064	0
employment_4<=X<7	0.934	0.368	0.263	0
employment_<1	-0.032	-0.171	-0.236	0
employment_>=7	0.345	0	0.090	0
employment_unemployed	-0.021	0	-0.054	0
personal_status_female div/dep/mar	0.213	0	-0.130	0
personal_status_male div/sep	-0.011	0	-0.130	0
personal_status_male mar/wid	0.625	0.101	0.099	0
personal_status_male single	0.609	0.168	0.161	0
other_parties_co applicant	0.072	0	-0.072	0
other_parties_guarantor	1.074	0.099	0.148	0
other_parties_none	0.29	0	-0.076	0
property_magnitude_life insurance	0.156	0	-0.119	0
property_magnitude_no known property	-0.031	-0.002	-0.008	0
property_magnitude_real estate	0.85	0.456	0.291	0
property_magnitude_car	0.461	0	-0.005	0
other_payment_plans_bank	0.392	0	-0.12	0
other_payment_plans_none	0.984	0.545	0.307	0
other_payment_plans_store	0.061	0	-0.187	0
housing_for free	0.735	0	-0.004	0
housing_own	0.65	0.376	0.188	0
housing_rent	0.052	0	-0.184	0
job_high qualif/self emp/mgmt	0.593	0	0.107	0
job_unemp/unskilled non res	0.422	0	0.-0.034	0
job_unskilled resident	0.223	0	0.-0.021	0
job_skilled	0.199	0	-0.052	0
own_telephone_none	0.574	0	-0.133	0
own_telephone_yes	0.853	0.151	0.133	0
foreign_worker_no	1.108	0	0.118	0
foreign_worker_yes	0.329	0	-0.118	0

Tabela 1: Tabela z pięcioma zestawami wartości dla różnych cech.