Analiza danych jakościowych

Łukasz Obrzut

Abstrakt

Raport dotyczy problemu klasyfikacji związanego z oczekiwaną długością życia pooperacyjnego u pacjentów chorych na raka płuca. Celem analizy jest zbudowanie modelu do przewidywania przynależności pacjenta do jednej z dwóch klas: klasa 1 – zgon w ciągu roku od zabiegu, klasa 2 – przeżycie powyżej jednego roku. Predyktory wykorzystywane w analizie pochodzą z okresu przedoperacyjnego. Zbiór danych charakteryzuje się niezbalansowaną strukturą, co wymaga zastosowania odpowiednich metryk oceny modelu. W ramach projektu zostanie zbudowany model predykcyjny oparty na regresji logistycznej, wykorzystujący metody selekcji forward oraz backward. Jakość modelu będzie oceniana za pomocą miary Gmean, która sprawdza, jak dobrze model równoważy czułość i swoistość, co jest ważne przy nierównomiernych danych.

Opis danych, sposób gromadzenie, źródło

Wykorzystane dane są dostępne pod linkiem: *https://archive.ics.uci.edu/dataset/277/thoracic+surgery+data*. Dane zostały zebrane retrospektywnie w Centrum Chirurgii Klatki Piersiowej we Wrocławiu od ponad 1200 kolejnych pacjentów, którzy przeszli duże resekcje płuc w związku z pierwotnym rakiem płuca w latach 2007–2011. Centrum jest związane z Kliniką Chirurgii Klatki Piersiowej Uniwersytetu Medycznego we Wrocławiu oraz Dolnośląskim Centrum Chorób Płuc, a baza danych badawczych stanowi część Krajowego Rejestru Raka Płuca, zarządzanego przez Instytut Gruźlicy i Chorób Płuc w Warszawie. Główny zbiór danych zawierał 139 predyktorów, z czego 36 pochodziło z okresu przedoperacyjnego, 37 z okresu okołooperacyjnego, a 46 z okresu pooperacyjnego. Ponadto, po wyeliminowaniu przykładów z brakującymi wartościami stworzono ostateczny zbiór danych zawierający 470 przykładów. Jednak na potrzeby naszego zadania rozważę 16 predyktorów, które zostały wybrane na podstawie zysku informacji z 36 predyktorów z okresu przedoperacyjnego. Tabela 1 zawiera opisy zmiennych użytych w badaniu, przedstawiając zarówno ich oryginalne nazwy bazowe, jak i zmienione nazwy. Zmiana ta została wprowadzona, aby ułatwić odniesienie się do danych oraz zapewnić większą intuicyjność w analizie. Warto zaznaczyć, że prawie wszystkie zmienne w zestawie danych są zmiennymi binarnymi. Ponadto są także zmienne uporządkowane tj.

- Wielkość guza wyrażona w czterech stopniach: OC11, OC14, OC12, OC13, gdzie OC11 oznacza najmniejszego guza, a OC14 największego.
- Skala Zubroda zawierająca 5 stopni, jednak w naszym zestawie uwzględniono jedynie 3 najlżejsze z nich, a konkretnie
  - 0 – sprawność prawidłowa, zdolność do samodzielnego wykonywania codziennych czynności.
  - 1 – obecność objawów choroby, możliwość chodzenia i wykonywania lekkiej pracy.
  - 2 – zdolność do wykonywania czynności osobistych, niezdolność do pracy, spędza w łóżku około połowy dnia.

  Być może wynika to z faktu dopuszczenia pacjenta do operacji.

Ponadto w zbiorze znajdują się trzy zmiennych: FVC, FEV1 (wartość podawana jest w litrach) oraz Wiek, które są zmiennymi ilościowymi.

Skrót	Zmieniona nazwa	Opis
DGN	DGN	Kody ICD-10 dla guzów pierwotnych, wtórnych i mnogich (7 poziomów).
PRE4	FVC	Całkowita ilość powietrza, którą można wydychać po głębokim wdechu.
PRE5	FEV1	Objetość powietrza wydmuchiwanego w ciągu 1 sekundy.
PRE6	Zubrod	Skala ECOG oceniająca sprawność pacjenta z chorobą nowotworową.
PRE7	Bol	Wskazuje obecność bólu przed operacją (True/False).
PRE8	Krwioplucie	Wskazuje obecność krwioplucia przed operacją (True/False).
PRE9	Duszność	Wskazuje obecność duszności przed operacją (True/False).
PRE10	Kaszel	Wskazuje obecność kaszlu przed operacją (True/False).
PRE11	Osłabienie	Wskazuje obecność osłabienia przed operacją (True/False).
PRE14	WielGuza	Skala TNM oceniająca wielkość pierwotnego guza.
PRE17	Cukrzyca2	Wskazuje obecność cukrzycy typu 2 przed operacją (True/False).
PRE19	ZawałMS	Określa, czy pacjent miał zawał serca w ostatnich 6 miesiącach.
PRE25	PAD	Określa, czy pacjent ma chorobę tętnic obwodowych.
PRE30	Palenie	Określa, czy pacjent pali papierosy (P – pali, K – nie pali).
PRE32	Astma	Określa, czy pacjent ma astmę (True/False).
AGE	Wiek	Określa wiek w chwili operacji
Risk1Yr	Ryzyko1R	Określa, czy pacjent umrze w ciągu roku po zabiegu (True - zgon).

Tabela 1. Skróty zmiennych i ich zmienione nazwy z opisami.

Przegląd danych

Podczas przeglądania danych napotkałem anomalię w zmiennej FEV1, dotyczącej objętości powietrza wydmuchiwanego w ciągu 1 sekundy. Wartości tej zmiennej nie odpowiadają rzeczywistym wartościom, ponieważ 15 obserwacji zawiera wartość powyżej 8 litrów, co jest nierealnym wynikiem z faktu, że pojemność płuc dorosłego mężczyzny wynosi średnio 5-6 litrów. Co więcej, 14 z tych obserwacji wskazuje wartości powyżej 60 litrów co może wynikać z błędu interpretacji pomiaru. Możliwe, że wartości te odnoszą się do procentu optymalnej wartości wydychanej FEV1, jednak brak takiej adnotacji w opisie danych. Z tego względu zdecydowałem się odrzucić te 15 przypadków z analizy zostawiając tym samym, w zbiorze danych 455 obserwacji, w tym 386 osób, które przeżyły rok po operacji oraz 69, które zmarły.

Podstawowe zależności między danymi

Dokładny test Fishera

Najpierw rozpatrzę swój zbiór danych pod kątem niezależności zmiennych. W tym celu zastosuję dokładny test Fishera, wykorzystujący symulacje Monte Carlo z parametrem B ustawionym na 20000. Zdecydowałem się na użycie tego testu, ponieważ zbiór danych jest stosunkowo mały, a w tabelach kontyngencji zdarzają się komórki zawierające mniej niż 5 obserwacji. Test ten nie jest jednak w stanie wykryć zależności monotonicznych, w związku z tym zastosuję współczynnik gamma Goodmana i Kruskala. Wyniki testu przedstawiono w Tabela 2.

	DGN	Zubrod	Bol	Krwioplucie	Duszność	Kaszel	Oslabienie	WielGuza	Cukrzyca2	ZawałMS	PAD	Palenie	Astma	Ryzyko1R
DGN	≈ 0	0.283	0.887	0.383	0.784	0.180	0.750	0.095	0.621	1	0.235	0.188	1	0.003
Zubrod	0.283	≈ 0	0.005	0.016	0.106	<0.001	<0.001	0.012	0.409	1	0.669	<0.000	1	0.097
Bol	0.89	0.005	≈ 0	<0.001	0.179	0.281	0.288	0.010	0.443	1	1	0.111	1	0.159
Krwioplucie	0.383	0.017	<0.001	≈ 0	0.063	0.181	0.145	0.535	0.797	1	0.082	0.718	1	0.088
Duszność	0.778	0.1	0.179	0.063	≈ 0	0.121	0.281	0.019	1	1	0.066	0.412	1.000	0.038
Kaszel	0.175	<0.001	0.281	0.181	0.121	≈ 0	0.006	0.014	0.7	1	1	<0.001	0.518	0.047
Oslabienie	0.751	<0.001	0.288	0.145	0.281	<0.001	≈ 0	0.266	0.153	0.313	0.631	0.013	1	0.083
WielGuza	0.098	0.013	0.009	0.527	0.019	0.014	0.267	≈ 0	0.262	1	1	0.892	1.000	0.004
Cukrzyca2	0.619	0.417	0.443	0.797	1	0.7	0.153	0.267	≈ 0	1	0.465	0.346	1	0.024
ZawałMS	1	1	1	1	1	1	0.314	1	1	≈ 0	1	1	1	1
PAD	0.235	0.658	1	0.082	0.063	1	0.631	1	0.465	1	≈ 0	0.361	1	0.348
Palenie	0.188	<0.001	0.111	0.718	0.412	<0.001	0.013	0.886	0.346	1	0.361	≈ 0	0.318	0.118
Astma	1	0.550	1	1	1	0.518	1	1	1	1	1	0.318	≈ 0	1
Ryzyko1R	0.003	0.097	0.159	0.088	0.038	0.047	0.083	0.004	0.024	1	0.348	0.118	1	≈ 0

Tabela 2. Tabela *p*-wartości dokładnego testu Fishera

Widzę, że dokładny test Fishera wykrył zależności między zmiennymi objaśniającymi, nie tylko między zmienną objaśnianą, a predyktorami. Niemniej jednak, zauważam podejrzane wyniki dla zmiennej *Astma* oraz *ZawałMS*. Dzieje się tak z powodu dużego niezbalansowania danych dla tych zmiennych. W obu przypadkach w naszej bazie danych znajduje się tylko 2 obserwacje, które mają wartość True. Uznaję te zmienne za mało przydatne w tworzeniu modelu, ponieważ ich skrajne niezbalansowanie sprawia, że nie dostarczają istotnych informacji statystycznych i decyduję się nie rozważać ich w dalszej analizie. Ponadto, dla zmiennych *PAD* i *Ryzyko1R* dokładny test Fishera również nie daje podstaw do odrzucenia hipotezy o niezależności zmiennych. W tym przypadku również występuje duże niezbalansowanie, ale nie jest ono tak skrajne, jak w przypadku zmiennych *Astma* i *ZawałMS*. Ponadto, zauważam, że test Fishera wykrywa również pewne zależności między zmienną *Zubrod*, a zmiennymi: *Bol, Krwioplucie*, *Duszność, Kaszel, Osłabienie, Palenie, WielGuza*. Intuicyjnie, zależności między skalą Zubroda, a wymienionymi zmiennymi wydają się mieć sens, ponieważ skala Zubroda pomaga określić stan ogólny i jakość życia pacjenta z chorobą nowotworową. Dodatkowo, istnieją zależności między zmienńś *Krwioplucie*, a *Bol, Kaszel i Osłabienie*, a także między *Kaszel*, a *Palenie*. Te zależności również wydają się mieć sens, gdyż związane są z ogólnym stanem zdrowia i zachowaniami pacjentów. Wszyskie zależności pokazane są w Tabeli 2.

Współczynnik gamma Goodmana i Kruskala

Aby sprawdzić, czy występują również zależności typu monotonicznego, użyję współczynnika gamma Kruskala, który jest miarą oceniającą takie zmiennych. Tabela 3 przedstawia wartości współczynnika gamma dla poszczególnych par zmiennych. Można zauważyć, że wysokie wartości współczynnika gamma występują dla par zmiennych, które w teście Fishera miały najniższe p-wartości. Sugeruje to, że dla tych par istnieją zależności monotoniczne.

Ponadto możemy zobaczyć w poniższej tabeli czerwone wartości, które oznaczają, że 95% przedział ufności nie zawiera zera. Widzimy tu między innymi, że jakaś zależność monotoniczna może istnieć pomiędzy diagnozą, a wielkością guza. Jednak co ciekawe dla zmiennych *WielGuza* oraz *Zubrod* mimo, że dla dokładnego testu Fishera widzieliśmy zależność tych zmiennych to 95% przedział ufności zawiera zero dla tych zmiennych. Możemy dostrzec inne monotoniczności, jednak musimy pamiętać o fakcie, że współczynnik gamma równy 0 nie musi implikować niezależności.

	DGN	Zubrod	Bol	Krwioplucie	Duszność	Kaszel	Oslabienie	WielGuza	Cukrzyca2	PAD	Palenie	Ryzyko1R
DGN	1	-0.0735	0.096	-0.250	-0.166	-0.144	-0.041	-0.221	0.184	-0.25	-0.123	0.077
Zubrod	-0.074	1	0.644	0.385	0.426	0.957	0.963	0.158	0.159	0.239	0.439	0.262
Bol	0.096	0.644	1	0.716	0.395	-0.234	-0.459	0.342	0.231	-1	-0.358	0.348
Krwioplucie	-0.251	0.385	0.716	1	0.452	0.231	0.251	0.149	0.048	0.596	-0.077	0.284
Duszność	-0.167	0.426	0.395	0.452	1	0.412	-0.423	0.434	-0.332	0.720	-0.214	0.48
Kaszel	-0.144	0.957	-0.234	0.231	0.412	1	0.647	0.286	0.108	0.140	0.483	0.306
Oslabienie	-0.041	0.963	-0.459	0.251	-0.423	0.647	1	-0.139	0.295	0.239	0.485	0.271
WielGuza	-0.221	0.158	0.342	0.149	0.434	0.286	-0.139	1.000	0.069	-0.084	0.094	0.345
Cukrzyca2	0.185	0.159	0.231	0.048	-0.332	0.108	0.295	0.069	1	0.284	-0.205	0.438
PAD	-0.25	0.239	-1	0.596	0.72	0.14	0.239	-0.084	0.284	1	1	0.308
Palenie	-0.124	0.439	-0.358	-0.077	-0.214	0.483	0.485	0.094	-0.205	1	1	0.340
Ryzyko1R	0.078	0.262	0.348	0.284	0.48	0.306	0.271	0.345	0.438	0.308	0.34	1

Tabela 3. Tabela wartości gamma Goodmana-Kruskala

Specyfikacje zastosowanych modeli i ich weryfikacje

W swoich rozważaniach będę korzystał z metod *forward* oraz *backward* doboru zmiennych do modelu regresji logistycznej.

- **Metoda forward** (selekcja do przodu) polega na rozpoczynaniu od pustego modelu i stopniowym dodawaniu zmiennych. Na każdym etapie do modelu dodawana jest ta zmienna, która najlepiej poprawia dopasowanie modelu.
- **Metoda backward** (selekcja do tyłu) zaczyna się od pełnego modelu zawierającego wszystkie zmienne. W kolejnych krokach usuwane są zmienne, które w najmniejszym stopniu przyczyniają się do dopasowania modelu.

Do wyboru odpowiedniego modelu regresji logistycznej w R użyję funkcji `step`, która implementuje zarówno selekcję backward, jak i forward, z minimalizowaniem kryterium AIC. Należy jednak pamiętać, że taki model nie zawsze musi być najlepszym modelem.

Jeśli chodzi o weryfikację, stosuję również wskaźnik Gmean, aby zapobiec sytuacji, w której model przewiduje każdy przypadek jako należący do klasy większościowej. Miara Gmean bierze ona pod uwagę zarówno czułość (TPR-True positive rate), jak i swoistość (TNR-True negative rate), zapewniając ocenę skuteczności modelu w odniesieniu do obu klas. Wyraża się ona wzorem:

$$\text{Gmean} = \sqrt{\text{TPR} \cdot \text{TNR}},$$

gdzie czułość (TPR) to stosunek wyników prawdziwie dodatnich do sumy prawdziwie dodatnich i fałszywie ujemnych, a swoistość jest stosunkiem wyników prawdziwie ujemnych do sumy prawdziwie ujemnych i fałszywie dodatnich. W naszym przypadku

- **Czułość (TPR):** stosunek liczby osób, które faktycznie zmarły w ciągu roku od operacji i model przewidział zgon, do liczby wszystkich osób, które faktycznie zmarły $(TPR = \frac{\text{TP}}{\text{TP+FN}})$.
- **Swoistość (TNR):** stosunek liczby osób, które przeżyły ponad rok po operacji i model przewidział przeżycie, do liczby wszystkich osób, które faktycznie przeżyły $(TNR = \frac{\text{TN}}{\text{TN+FP}})$.

Oczywiście wartość współczynnika należe do przedziału $[0, 1]$, gdzie czym większa wartość współczynnika tym model wydaje się być lepszy.

Modele regresji logistycznej

Podział na zbiór treningowy i testowy

Z powodu ograniczonej liczby danych, zdecydowałem się podzielić zbiór danych na 85% danych treningowych i 15% danych testowych. Ponadto, aby zachować równowagę klas w obu zbiorach, użyłem funkcji *createDataPartition* z bibloteki *caret*. Zanim przystąpię do budowania modeli, warto zauważyć, że z powodu niezbalansowania danych lepsze wyniki będą otrzymywane dla progów mniejszych niż 0.5. Dla progu 0.5 prezentowane modele będą wykazywać tendencję do kategoryzowania większości przypadków do jednej klasy.

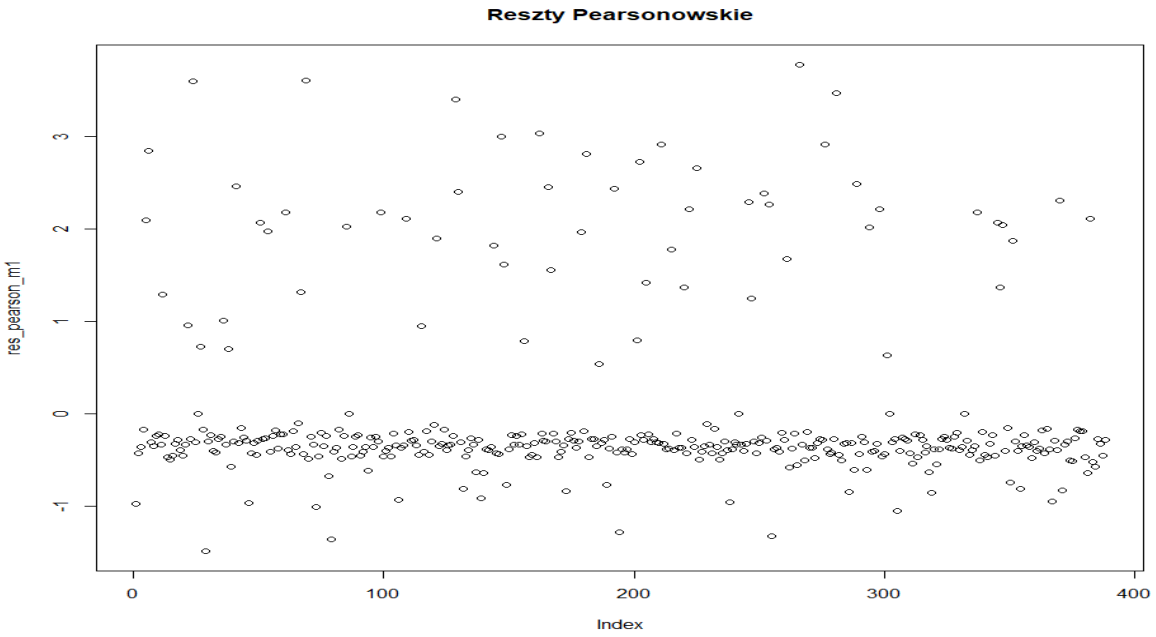
Model nr 1

Do pierwszego modelu zastosowałem selekcję backward, nie uwzględniając interakcji między zmiennymi. Dla danych treningowych uzyskałem AIC na poziomie 316.99. Taki sam dobór zmiennych otrzymałem również przy użyciu selekcji forward. Model ten wykorzystałem do oceny współczynnika Gmean. Gmean osiąga najwyższą wartość dla progu 0.165383 i wynosi około 0.7. Widoczne są także małe *p*-wartości dla testu istotności typu Walda dla zmiennej DGN.

Z	Wartości estymowane	Błąd standardowy	statystyka z	Pr($>\|z\|$)	
(Intercept)	-15.62370	2399.54486	-0.007	0.9948	
DGN2	14.59013	2399.54476	0.006	0.9951	
DGN3	14.41250	2399.54486	0.006	0.9952	
DGN4	14.74168	2399.54476	0.006	0.9951	
DGN5	16.91899	2399.54486	0.007	0.9949	
DGN6	0.41790	2668.98540	0.000	0.9999	
DGN8	18.11434	2399.54519	0.008	0.9940	
FEV1	-0.51787	0.22265	-2.326	0.0200*	
BólTRUE	0.97409	0.52524	1.855	0.0637	.
PalenieTRUE	0.97029	0.50217	1.931	0.0535	.
DusznośćTRUE	0.96841	0.60407	1.603	0.1089	
WielGuza.L	1.37653	0.43373	3.174	0.0015	**
WielGuza.Q	0.40065	0.43601	0.918	0.3581	
WielGuza.C	0.06782	0.43350	0.156	0.8757	

Tabela 4. Podsumowanie modelu nr 1.

Możemy również zobaczyć Rysunek 1, gdzie widoczne jest zachowywanie reszt Pearsonowskich w modelu (ze względu podobieństwa zachowania reszt Personowskich z każdym z modeli umieszczam tylko rysunek dla modelu nr 1). Reszty Pearsonowskie pokazują, że w modelu widoczne jest niedopasowanie modelu do części obserwacji.



Rysunek 1. Reszty Pearsonowskie dla modelu 1

Model 2

Następnym krokiem w budowie modelu było rozszerzenie modelu pełnego o dwie interakcje: Palenie: WielGuza oraz WielGuza:Krwioplucie. Następnie, przy użyciu metody selekcji forward, dobrałem optymalny zestaw zmiennych dla drugiego modelu. Okazało się, że wybrano zmienne zawarte w modelu nr 1 oraz interakcję Palenie : WielGuza. Spowodowało to spadek wartości AIC do 307.3872.

Wskaźnik Gmean osiągnął swoją najwyższą wartość w pobliżu 71 przy progu klasyfikacji wynoszącym 0.165. Rysunek 2 przedstawia zależność wartości Gmean od różnych progów klasyfikacji dla danych treningowych.

Model 3

Ostatnim podejściem będzie zbudowanie modelu selekcją forward gdy z modelu pełnego usunę zmienną DGN. Wyniki znajdują się w poniższej tabeli. Ponadto Gmean dla tego modelu zachowuje się podobnie

Zmienne	Wartości estymowane	Błąd standardowy	statystyka z	Pr(>|z|)
(Intercept)	-0.8894852	0.6770286	-1.314	0.18891
WielGuza.L	1.2819321	0.4057493	3.159	0.00158 **
WielGuza.Q	0.3208294	0.4219235	0.760	0.44702
WielGuza.C	-0.0006659	0.4168068	-0.002	0.99873
FEV1	-0.4531052	0.2125420	-2.132	0.03302 *
BolTRUE	0.8883814	0.5076097	1.750	0.08010 .
PalenieTRUE	0.6558729	0.4417677	1.485	0.13764
DusznośćTRUE	0.9062098	0.5854446	1.548	0.12165

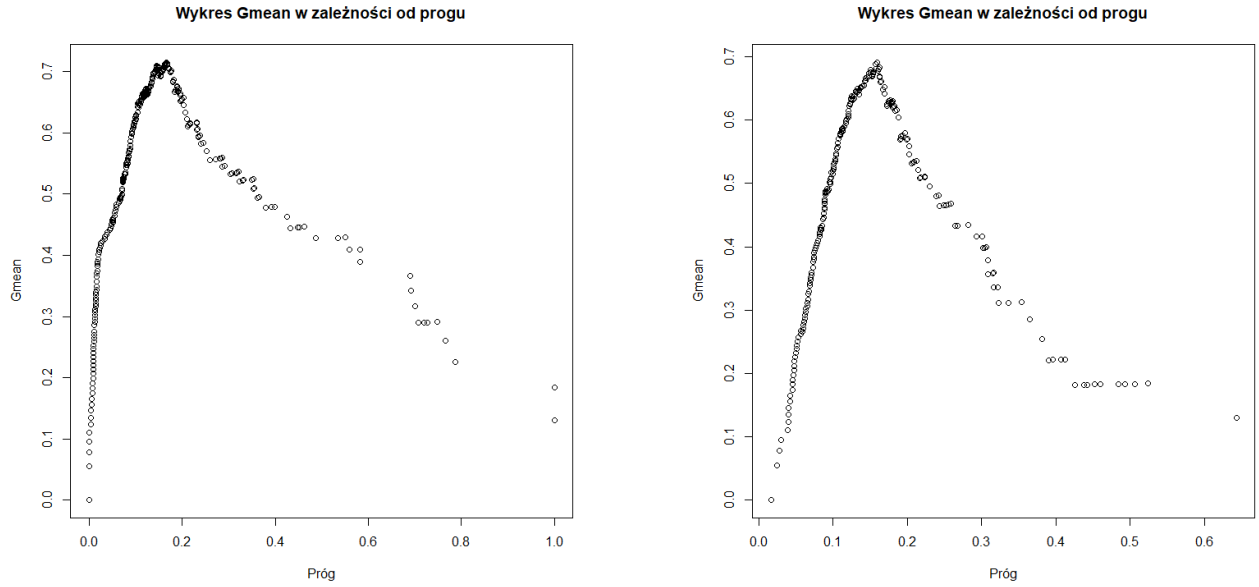
Tabela 5. Wyniki dla modelu nr 3.

jak w przypadku modelu 1 i 2. Jednakże wartość Akaike oraz dewiancja w modelu jest wyższa niż w

poprzednich modelach. Ze względu na podobieństwo wartości Gmean dla wszystkich modeli swój wybór oprę na teście ilorazu wiarygodności dla zagnieżdzonych modeli (w naszym przypadku modele są w sobie zagnieżdzone). Przy użyciu funkcji *anova* z dodatkowym argumentem $test =$ "*Chisq*" porównuję modele otrzymując:

- porównując model 3 z modelem 1 p-wartość wynosi 0.003420
- porównując model 3 z modelem 2 p-wartość wynosi 0.00136

Z powyższych porównań widzę, że na poziomie istotności 5% model nr 2 wydaje się najlepszy w kontekście rozważanych modeli i testu.



Rysunek 2. Gmean w zależności od przyjętego progu w modelu 2 (po lewej) i modelu 3 (po prawej), dla modelu 1 podobne rezultaty

Sprawdzenie modelu na zbiorze testowym

Jako finalny model do predykcji wybieram model nr 2 oraz próg ustawiam na wartość 0.16 na podstawie Rysunku 2. Zatem dla tego modelu sprawdzam wyniki. Okazuje się, że wskaźnik Gmean dla danych testowych wyniósł w pobliżu 0.592. Przewidywania tego modelu ze względu m.in na poziom niezbalansowania zmiennej objaśniającej jest mocno oparty na doborze odpowiedniego progu. Dzięki temu doborowi wzrastają zdolności predykcyjne modelu. Ponadto dla danych testowych w Tabeli 6 pokazana jest macierz błędów.

Przewidziana/Obserwowana	**Przeżycie**	**Śmierć**
Przewidziane przeżycie	40	5
Przewidziana śmierć	17	5

Tabela 6. Macierz błędu dla danych testowych (model 2)

Można zauważyć, że model, dzięki odpowiednim zmiennym i parametrze progu wykazuje pozytywne zdolności predykcyjne

Podsumowanie

Na podstawie pierwszej części projektu zaobserwowałem zależności między różnymi zmiennymi, które mogą być istotne w okresie przedoperacyjnym. Niestety, z powodu ograniczonej liczby danych, nie mogłem uwzględnić wystarczającej liczby przypadków dotyczących wpływu zawału mięśnia sercowego czy astmy na ryzyko śmierci po zabiegu operacji raka płuc. Większy zbiór danych mógłby dostarczyć istotnych informacji na ten temat.

Jeśli chodzi o wnioski związane z budową modelu predykcyjnego, warto rozważyć zastosowanie innego podejścia. Pomimo niewielkiej liczby danych i nierównomiernego rozkładu zmiennej objaśnianej, udało się stworzyć model, który osiągnął wynik Gmean na poziomie 59%. Aby lepiej zrozumieć możliwości analizy tego zbioru danych oraz zastosowanych na nim modeli, warto spojrzeć na źródło nr 1.

Odwołania

[1] M. Zięba, J. M. Tomczak, M. Lubicz, J. Świątek, *Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients*, Applied Soft Computing, vol.14, 2014, 99-108

[2] *Thoracic Surgery Data* , UCI Machine Learning Repository