Tematy projektu: Determinanty zaufania wśród brytyjczyków- regresja logistyczna

Autor: Sebastian Boruch

Spis treści

1. Cel badania i opis zbioru

Prezentowany projekt stanowi analizę zaufania wśród brytyjskiego społeczeństwa. W tym projekcie do analizy badanego problemu wykorzystano regresję uporządkowaną i binarną (forward i backward).

Celem przeprowadzonej analizy było zidentyfikowanie w jakim stopniu czynniki społeczno - demograficzne wpływają na osobiste odczucie dotyczące tego, czy ludzie chcą badanego wykorzystać, czy też są wobec niego uczciwi. Analiza umożliwi również zbadanie jak bardzo brytyjskie społeczeństwo ufa sobie nawzajem. W budowie modeli regresji logistycznej wykorzystano dane pochodzące z European Social Survey. Jest to platforma gromadząca dane dotyczące wzorców zachowań, nastawień oraz ocen kluczowych sfer życia społecznego wielu krajów Europy.

W projekcie wykorzystano dane z badania przeprowadzonego w Wielkiej Brytanii w 2016 roku. Oryginalny zbiór zebrany na podstawie ankiet przeprowadzanych osobiście liczy 1959 obserwacji i opisuje odpowiedzi na 499 pytań. Autor do przeprowadzenia tego raportu wybrał 11 zmiennych. Te pytania wraz z nazwami zmiennych do nich przypisanych znajdują się w tabeli poniżej:

Tabela 1. Charakterystyki zmiennych

Pytanie	**Nazwa zmiennej**	**Typ**	**Wartości**
Większość ludzi chce cię wykorzystać czy są wobec ciebie szczerzy?	PPLFAIR	Kategoryczna porządkowa – zmienna celu	1 – nieufny 2 – średnioufny 3 – ufny
Ile masz lat?	AGEA	Numeryczna	15-94
Ile ukończyłaś/łeś lat edukacji?	EDUYRS	Numeryczna	0-54
W podanej skali, jak bardzo zgadzasz się ze stwierdzeniem, że integracja europejska powinna się pogłębiać?	EUFTF	Nominalna (liczby całkowite od 0 do 10)	0 – integracja postąpiła za daleko 10 – należy pogłębiać integrację
W podanej skali, jak bardzo jesteś szczęśliwa/wy?	HAPPY	Nominalna (liczby całkowite od 0 do 10)	0 – bardzo nieszczęśliwa/wy 10 – bardzo szczęśliwa/wy
Jak oceniasz swoje zdrowie?	HEALTH	Nominalna (liczby całkowite od 1 do 5)	1 – bardzo dobrze 5 – bardzo źle
W podanej skali, jak ważne jest dla ciebie odnoszenie sukcesów i bycie docenionym?	IPSUCES	Nominalna (liczby całkowite od 1 do 6)	1 – jest dla mnie bardzo ważne 6 – nie jest dla mnie ważne
W podanej skali, jak bardzo jesteś zadowolona/ny z funkcjonowania	STFDEM	Nominalna (liczby całkowite od 0 do 10)	0 – bardzo niezadowolona/ny 10 – bardzo zadowolona/ny

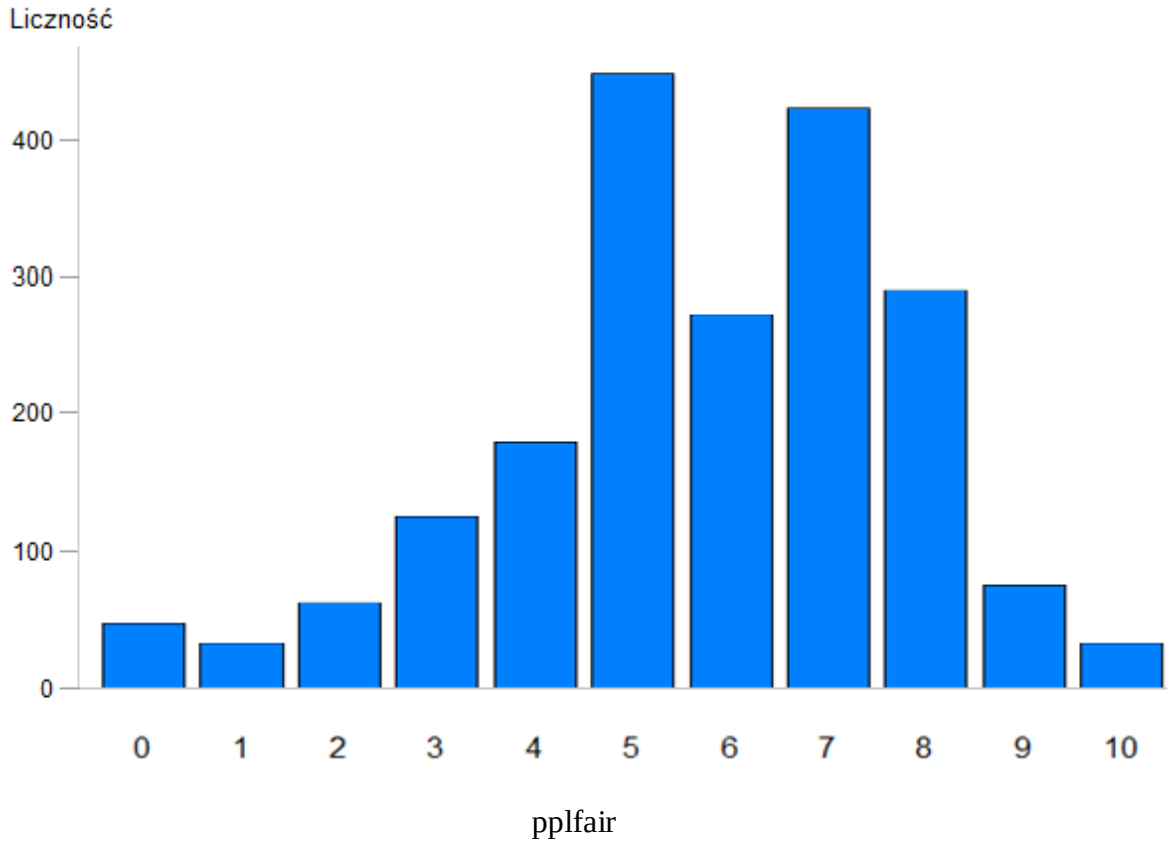
demokracji w Twoim kraju?			
Jaki jest ogólny przychód netto gospodarstwa domowego?	HINCTNTA	Nominalna (liczby całkowite od 1 do 10 oznaczające decyle w społeczeństwie)	1 – pierwszy decyl 10 – dziesiąty decyl
Jak bardzo ufasz parlamentowi w twoim kraju?	TRSTPRL	Nominalna (liczby całkowite od 0 do 10)	0 – nie ufam 10 – ufam całkowicie
Ilu masz podwładnych w pracy?	NJBSPV	Numeryczna	0-50

Źródło: Opracowanie własne na podstawie danych z European Social Survey 2016

2. Statystyki opisowe zmiennych

Analizowana zmienna celu PPLFAIR wyjściowo miała 10 kategorii - od 0 do 9, gdzie 0 określało największy poziom nieufności i poczucia wykorzystania przez społeczeństwo. Poniżej na Wykresie 1, zilustrowano początkowy rozkład zmiennej. Najpopularniejszą kategorią wśród ankietowanych z UK okazał się średni poziom równy 6.

Rysunek 1. Rozkład zmiennej PPLFAIR - 11 poziomów



Źródło: opracowanie własne

Ten rozkład zmiennej okazuje się być trudny do zinterpretowania poprzez regresję logistyczną. Sam zbiór cechował się również dużą ilością braków danych (ponad 9% na całym zbiorze). Dlatego autor raportu zdecydował, aby brakujące dane lub odpowiedzi oznaczone jako ,,brak odpowiedzi", ,,nie dotyczy", ,,nie wiem" lub ,,odmowa odpowiedzi" zastąpić danymi wygenerowanymi przez metodę średniej ruchomej (PROC MI w SAS). Metoda średniej ruchomej jest metodą imputacji dostępną m.in. dla zmiennych ciągłych. Jest ona podobna do metody regresji, z tą różnicą, że dla każdej brakującej wartości przypisuje wartość losowo z zestawu obserwowanych wartości, których przewidywane wartości są najbliższe przewidywanej wartości dla brakującej wartości z symulowanego modelu regresji. Metoda średniej ruchomej zapewnia, że wartości imputowane są wiarygodne i mogą być bardziej odpowiednie niż metoda regresji, jeśli założenie normalności zostanie naruszone. Zastosowanie tej metody umożliwiło użycie wszystkich obserwacji do każdego z modeli regresji.

Na potrzeby omawianej regresji uporządkowanej skategoryzowano zmienną PPLFAIR na 3 kategorie:

1 – nieufny

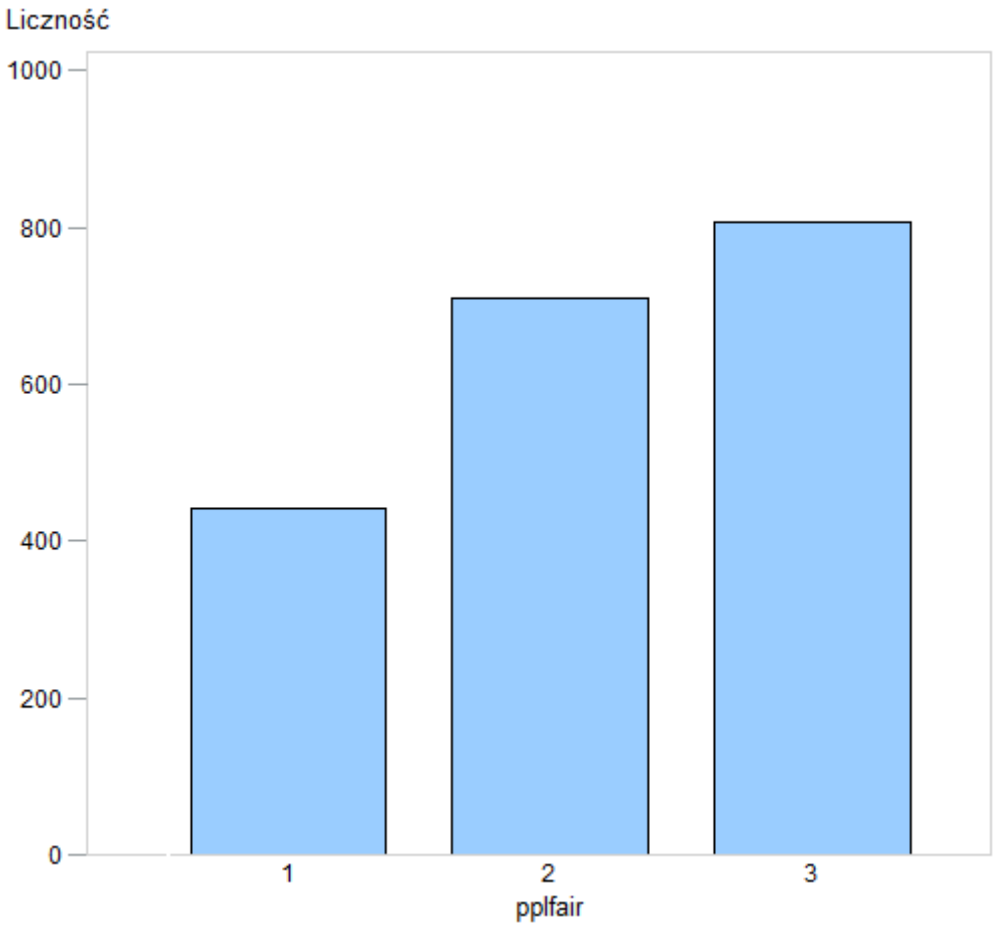
2 – średnioufny

3 – ufny.

W pierwszej kategorii znalazły się osoby, które oceniając uczciwość polskiego społeczeństwa wybrały ze skali od 0 do 4. Kolejna kategoria skala 5 i 6, zaś ostatnia zawierała odpowiedzi ze skali od 7 do 10.

Zmienna objaśniana jest zmienną porządkową rosnącą, poziom 3 określa najwyższy poziom ufności z zaprezentowanych. Na Rysunku 2 przedstawiono histogram zmiennej PPLFAIR po modyfikacji.

Rysunek 2. Rozkład zmiennej PPLFAIR - 3 poziomy



Źródło: opracowanie własne

Histogram nie prezentuje zbilansowanej liczby dla każdej kategorii. Jednak, dzięki zastąpieniu braków danych przewidzianymi wartościami, liczność każdej z kategorii przekracza 400, co umożliwia przeprowadzenie wiarygodnej analizy zbioru danych. W celu dokładnego przeanalizowania zmiennej celu, przygotowano tabelę z liczebnością oraz procentem występowania każdej z kategorii.

Tabela 2. Rozkład zmiennej PPLFAIR

pplfair	Liczebność	Procent	Liczebność skumulowana	Procent skumulowany
1	441	22.51	441	22.51
2	710	36.24	1151	58.75
3	808	41.25	1959	100.00

Źródło: opracowanie własne

Kategoria 1 stanowi 22,5% odpowiedzi, kategoria 2 stanowi 36% a kategoria 3 około 42%.

Statystyki zmiennych objaśniających

W dalszych rozdziałach raportu ukazane będzie, iż większość ze zmiennych jest istotna statystycznie. Jednym z założeń analizy regresji jest brak występowania współliniowości zmiennych objaśniających). Wprowadzając do modelu regresji silnie skorelowane ze sobą zmienne wprowadzamy do modelu (przy każdej zmiennej) małą bądź zerową unikalną "część wyjaśnienia" zmiennej zależnej. W zależności od sposobu liczenia jeden z predyktorów silnie powiązanych straci swoją "moc" przewidywania.

Współczynnik korelacji Pearsona służy do sprawdzenia czy dwie zmienne ilościowe są powiązane ze sobą związkiem liniowym. Wynik Pearsona może wahać się od -1 do 1. Wartości skrajne czyli -1 i 1 oznaczają idealną, totalną korelację między zmienną A i zmienną B. Wynik równy ‚zero' oznacza brak współwystępowania wartości tych dwóch zmiennych w naturze (brak korelacji).

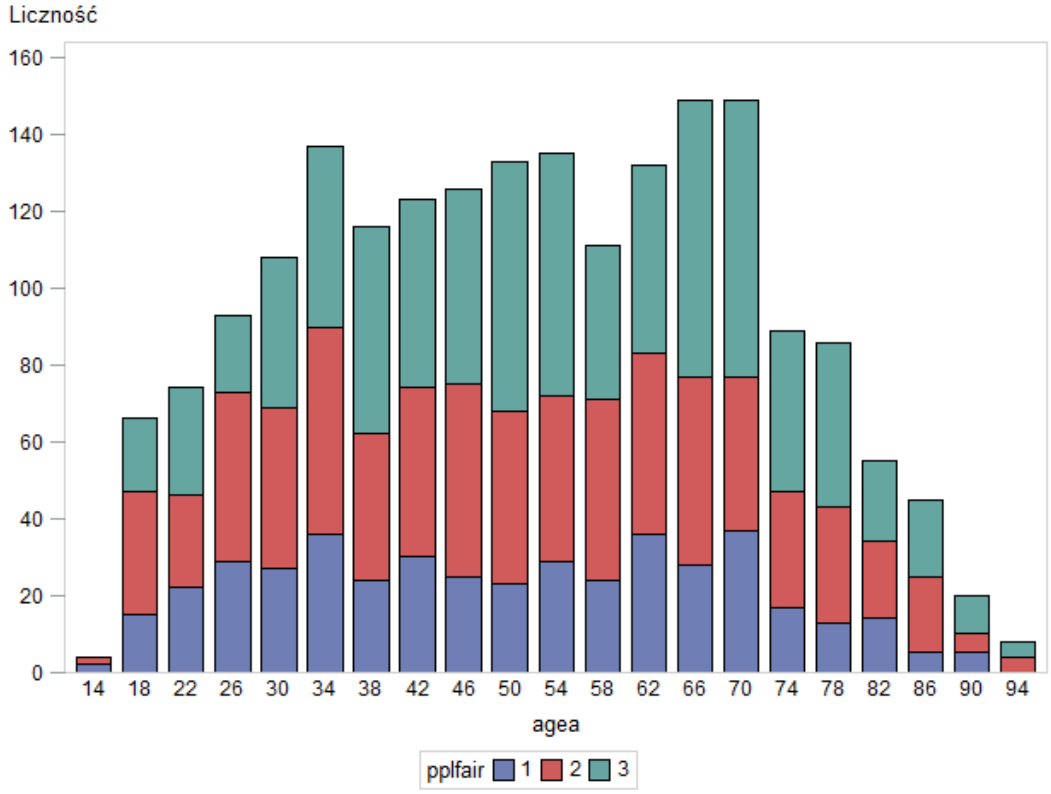
Tabela 3. Współczynniki korelacji Pearsona

Współczynniki korelacji Pearsona, N = 1959 Prawd. > |r| przy H0: rho=0											
	pplfair	agea	eduyrs	euftf	happy	health	ipsuces	stfdem	hinctnta	trstprl	njbspv
pplfair	1.00000	0.09309	0.12633	0.09068	0.19302	-0.19863	0.04564	0.17257	0.16454	0.27069	0.07893
		<.0001	<.0001	<.0001	<.0001	<.0001	0.0434	<.0001	<.0001	<.0001	0.0005
agea	0.09309	1.00000	-0.18991	-0.19139	0.03771	0.23335	0.27991	0.01356	-0.17371	-0.00569	0.17036
	<.0001		<.0001	<.0001	0.0952	<.0001	<.0001	0.5486	<.0001	0.8015	<.0001
eduyrs	0.12633	-0.18991	1.00000	0.09662	0.06994	-0.21209	-0.08042	0.07602	0.37423	0.16824	-0.02751
	<.0001	<.0001		<.0001	0.0020	<.0001	0.0004	0.0008	<.0001	<.0001	0.2236
euftf	0.09068	-0.19139	0.09662	1.00000	0.04473	-0.07975	-0.06648	0.03326	-0.01873	0.09955	-0.14235
	<.0001	<.0001	<.0001		0.0478	0.0004	0.0032	0.1411	0.4073	<.0001	<.0001
happy	0.19302	0.03771	0.06994	0.04473	1.00000	-0.33566	-0.00988	0.26550	0.17761	0.23403	-0.01065
	<.0001	0.0952	0.0020	0.0478		<.0001	0.6622	<.0001	<.0001	<.0001	0.6376
health	-0.19863	0.23335	-0.21209	-0.07975	-0.33566	1.00000	0.06367	-0.19932	-0.30138	-0.20990	0.11185
	<.0001	<.0001	<.0001	0.0004	<.0001		0.0048	<.0001	<.0001	<.0001	<.0001
ipsuces	0.04564	0.27991	-0.08042	-0.06648	-0.00988	0.06367	1.00000	-0.07797	-0.06729	-0.10227	0.04612
	0.0434	<.0001	0.0004	0.0032	0.6622	0.0048		0.0006	0.0029	<.0001	0.0413
stfdem	0.17257	0.01356	0.07602	0.03326	0.26550	-0.19932	-0.07797	1.00000	0.11395	0.54521	0.02471
	<.0001	0.5486	0.0008	0.1411	<.0001	<.0001	0.0006		<.0001	<.0001	0.2744
hinctnta	0.16454	-0.17371	0.37423	-0.01873	0.17761	-0.30138	-0.06729	0.11395	1.00000	0.15220	0.08184
	<.0001	<.0001	<.0001	0.4073	<.0001	<.0001	0.0029	<.0001		<.0001	0.0003
trstprl	0.27069	-0.00569	0.16824	0.09955	0.23403	-0.20990	-0.10227	0.54521	0.15220	1.00000	-0.04427
	<.0001	0.8015	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		0.0501
njbspv	0.07893	0.17036	-0.02751	-0.14235	-0.01065	0.11185	0.04612	0.02471	0.08184	-0.04427	1.00000
	0.0005	<.0001	0.2236	<.0001	0.6376	<.0001	0.0413	0.2744	0.0003	0.0501	

Źródło: opracowanie własne

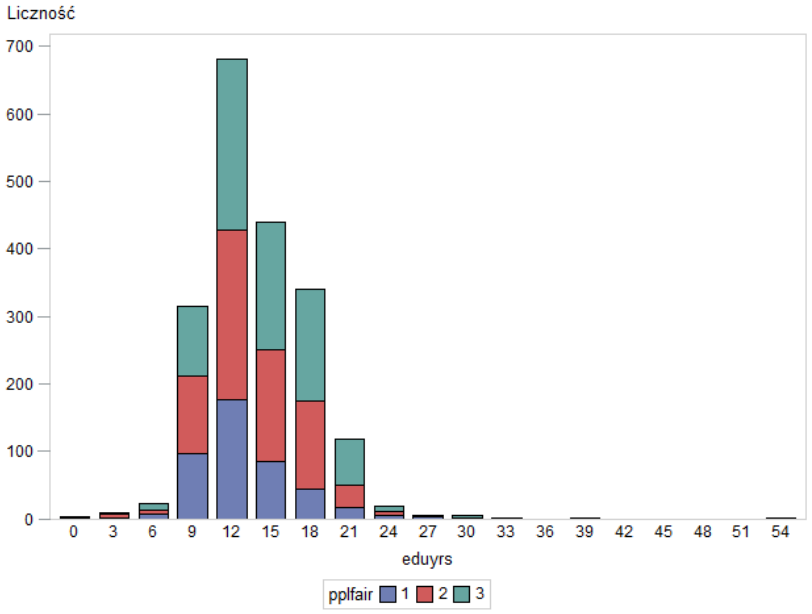
Jak widać po wynikach tabeli, prawie wszystkie zmienne są skorelowane ze sobą w stopniu słabym bądź bardzo słabym, co pozwala na przeprowadzenie analizy regresji. Korelacją silną jest tylko ta pomiędzy stfdem a trstprl, ale zmienna stfdem nie została uwzględniona w wynikach analizy.

Rysunek 3. Rozkład zmiennej AGEA z uwzględnieniem PPLFAIR



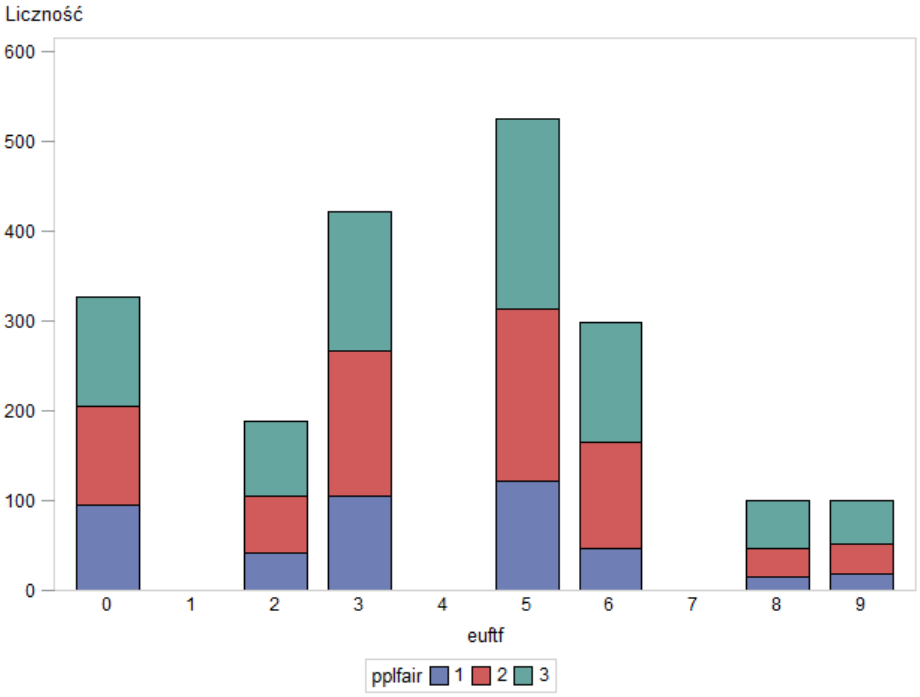
Źródło: opracowanie własne

Rysunek 4. Rozkład zmiennej EDUYRS z uwzględnieniem PPLFAIR



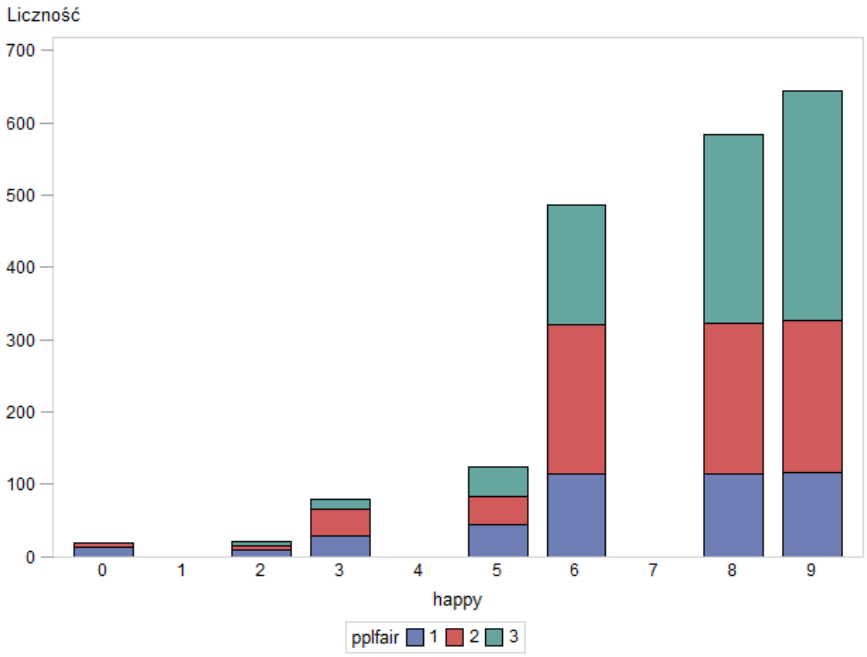
Źródło: opracowanie własne

Rysunek 5. Rozkład zmiennej EUFTF z uwzględnieniem PPLFAIR



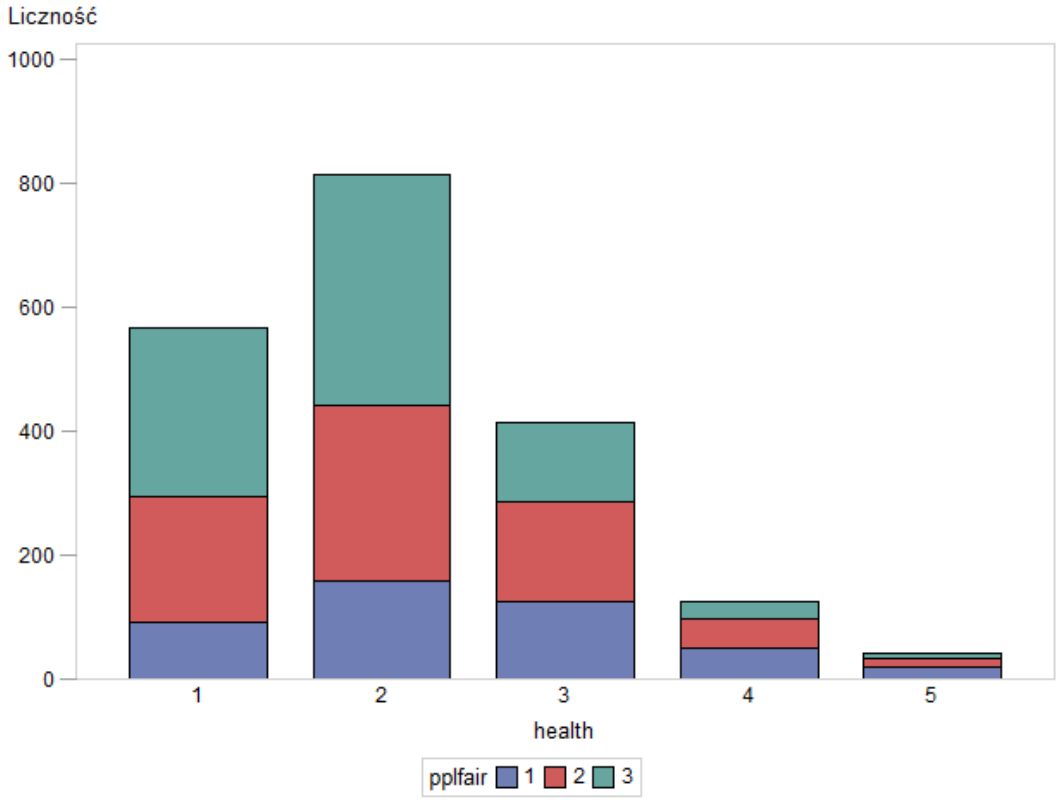
Źródło: opracowanie własne

Rysunek 6. Rozkład zmiennej HAPPY z uwzględnieniem PPLFAIR



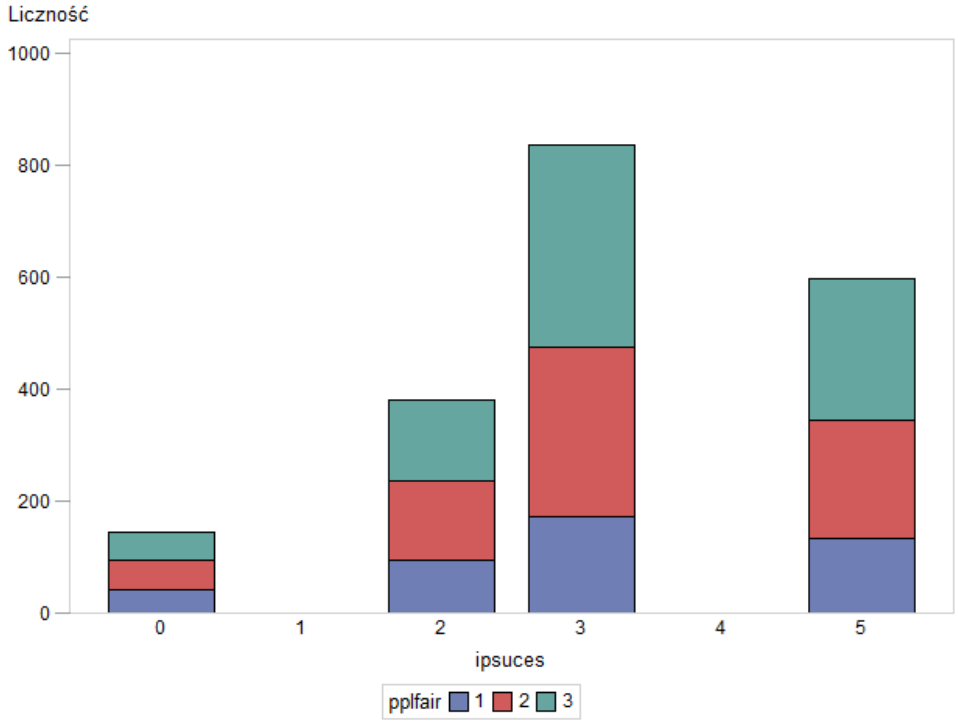
Źródło: opracowanie własne

Rysunek 7. Rozkład zmiennej HEALTH z uwzględnieniem PPLFAIR



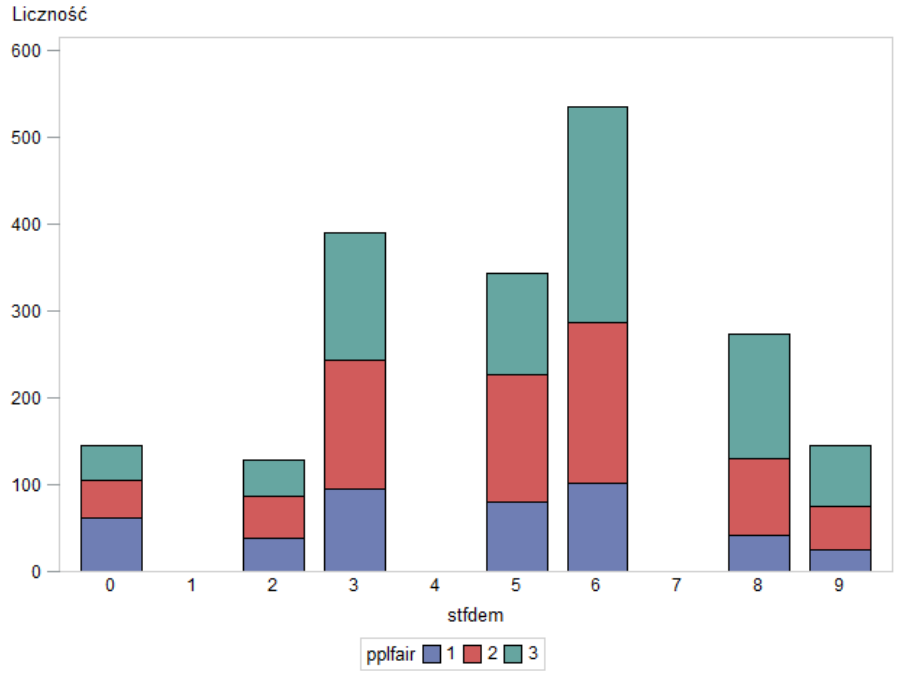
Źródło: opracowanie własne

Rysunek 8. Rozkład zmiennej IPSUCES z uwzględnieniem PPLFAIR



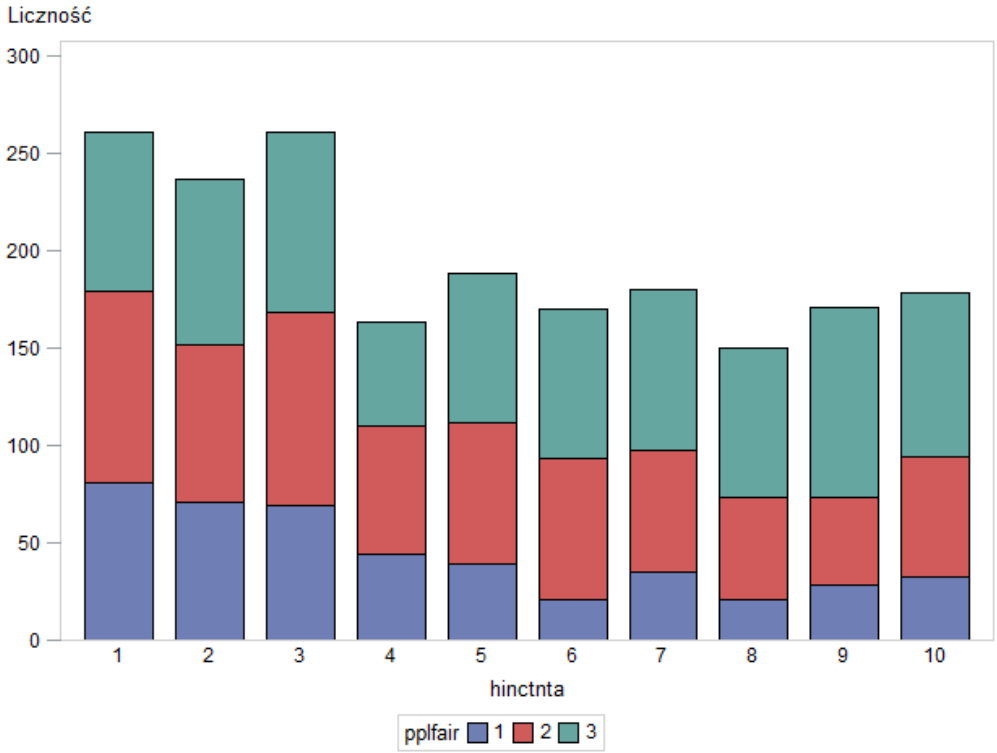
Źródło: opracowanie własne

Rysunek 9. Rozkład zmiennej STFDEM z uwzględnieniem PPLFAIR



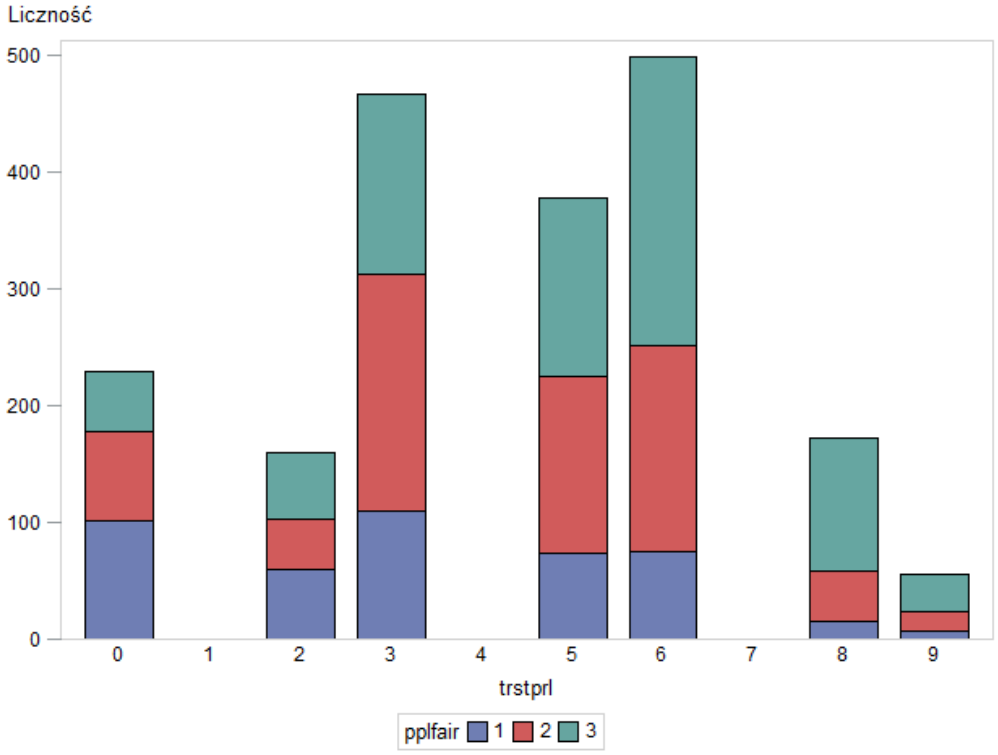
Źródło: opracowanie własne

Rysunek 10. Rozkład zmiennej HINCTNTA z uwzględnieniem PPLFAIR



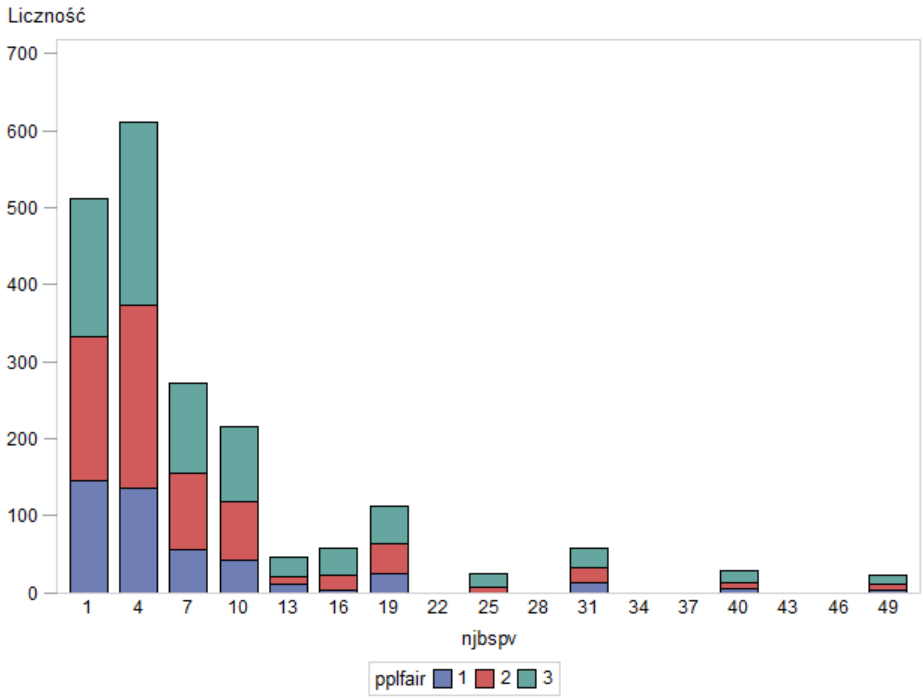
Źródło: opracowanie własne

Rysunek 11. Rozkład zmiennej TRSTPRL z uwzględnieniem PPLFAIR



Źródło: opracowanie własne

Rysunek 12. Rozkład zmiennej NJBSPV



Źródło: opracowanie własne

3. Model uporządkowanej regresji logistycznej

W przedstawionym projekcie zaprezentowano model uporządkowanej regresji logistycznej, gdzie zmienna objaśniania jest zmienną jakościową Y przyjmującą 3 kategorie, których wartość uzależniona jest od poziomu zmiennych niezależnych X1,X2,...,Xk (jakościowych bądź ilościowych). Podstawowym założeniem modelu uporządkowanej regresji logistycznej jest założenie o proporcjonalności odds. Przy spełnionym założeniu proporcjonalności odds, iloraz szans (odds ratio) mierzący efekt wpływu wystawienia na ryzyko dla szacowanych wariantów porównań będzie równy, niezależnie od tego, w jaki sposób zostaną pogrupowane kategorie uporządkowania. W modelu regresji porządkowej nie interesuje nas, do którego z porównań dany parametr beta się odnosi- fakt ten jest konsekwencją założenia proporcjonalności odds. Gdyby założenie to nie zostało spełnione, alternatywnym wyjściem byłaby estymacja modelu o postaci wielomianu.

W przypadku zmiennych czynnikowych zastosowano kodowanie porządkowe. W tym celu dokonano odpowiednich zmian w kodzie programu. Dodatkowo do budowy modelu włączono interakcje zmiennych jakościowych, ilościowych i jakościowo-ilościowych. Żadna z interakcji nie okazała się statystycznie istotna.

Dla modelu za funkcję linkująca przyjęto LOGIT. Dla prezentowanego modelu z trzema kategoriami zmiennej celu możliwe są dwa umiejscowienia cut-point, czyli uzyskać można maksymalnie 2 różne ilorazy szans. Natomiast po przeanalizowaniu liczebności kategorii, jedynym sensownym rozwiązaniem okazuje się kategoryzacja zmiennej w następujący sposób:

$$\text{PPLFAIR} = \langle 1,2 \rangle \text{ vs } \text{PPLFAIR} = 3$$

Model został zbudowany w oparciu o krokową selekcję zmiennych. Jest to modyfikacja wyboru następnych zmiennych i różni się od metody forward i backward tym, że zmienne, które zostaną raz zaklasyfikowane do modelu nie muszą w nim pozostać. Proces selekcji krokowej kończy się, kiedy żadna ze zmiennych niezaklasyfikowanych do modelu nie jest istotna na poziomie wystarczającym do wstawienia do modelu a wszystkie zaklasyfikowane charakteryzują się poziomem istotności wystarczającym do pozostawienia ich w modelu. Dla prezentowanego modelu poziom istotności określono jako 0,05.

3.1. Wyniki modelu uporządkowanej regresji logistycznej

W tej części projektu zaprezentowano wyniki modelu uporządkowanej regresji logistycznej.

Tabela 4. Podstawowe wyniki modelu

Informacje o modelu	
Zbiór	WORK.TMPMOD
Zmienna objaśniana	__RESPONSE
Liczba poziomów odpowiedzi	3
Model	logit skumulowany
Technika optymalizacji	Ocena Fishera

Wczytano obserwacji	1959
Użyto obserwacji	1959

Profil odpowiedzi		
Wartość uporządkowana	__RESPONSE	Całkowita liczebność
1	01: 1	441
2	02: 2	710
3	03: 3	808

Źródło: Opracowanie własne

Każda z 1959 obserwacji została użyta w modelu dzięki wykorzystaniu PPM (patrz str. 5).

Tabela 5. Status zbieżności modelu

Status zbieżności
Kryterium zbieżności (GCONV=1E-8) spełnione.

Źródło: Opracowanie własne

Kryterium zbieżności zostało spełnione, a więc znaleziono parametry największej wiarygodności.

Kolejno zweryfikowano test oceny przy założeniu proporcjonalności szans. Spełnienie założenia o niezmienności odds ratio jest podstawowym warunkiem estymacji uporządkowanego modelu regresji liniowej. Do weryfikacji tego założenia wykorzystywany jest test punktowy z poniższymi hipotezami:

H0: założenie o proporcjonalności odds jest w mocy

H1: założenie o proporcjonalności odds nie jest w mocy.

Tabela 6. Test proporcjonalności szans

Test oceny przy założeniu proporcjonalności szans		
Chi-kwadrat	DF	Pr. > chi-kw.
54.4309	45	0.1583

Źródło: Opracowanie własne

Wartośc p- value dla testu oceny przy założeniu proporcjonalności szans przy ostatnim kroku wyniósł 0,1583, a więc brak podstaw po odrzucenia hipotezy zerowej mówiącej o proporcjonalności szans- założenie zostało spełnione. Szanse w różnych grupach są proporcjonalne, co oznacza, że bez względu na podział kategorii zmiennej objaśnianej (czyli wybór, które kategorie są traktowane jako referencyjne) oszacowania ilorazów szans pozostają takie same.

Następnym etapem analizy była weryfikacja istotności parametrów modelu testując globalną hipotezę zerową:

H0:βj=0 (wszystkie parametry są równe zero- badany zestaw zmiennych objaśniających jest nieodpowiedni), przy hipotezie alternatywnej:

H1:βj ≠ 0 (przynajmniej jeden parametr jest istotny statystycznie).

Tabela 7. Test Beta=0

Testowanie globalnej hipotezy zerowej: BETA=0			
Test	Chi-kwadrat	DF	Pr. > chi-kw.
Iloraz wiarygod.	349.5262	45	<.0001
Wynik punktowy	314.8947	45	<.0001
Wald	303.5683	45	<.0001

Źródło: Opracowanie własne

Wyniki testów, zarówno Walda jak i Score, przyjmują wartości p-value na poziomie niższym niż 0,001, co sugeruje odrzucenie hipotezy zerowej na rzecz alternatywnej, czyli przynajmniej jeden ze współczynników modelu jest istotnie różny od zera.

Dalej zbadano, czy każdy estymowany parametr pojedynczo istotnie różni się od zera. Metoda krokowej selekcji, pozwoliła pozostawić w modelu tylko zmienne statystycznie istotne. Kolejne kroki selekcji ukazuje Tabela 10.

Tabela 8. Wyniki selekcji krokowej

Podsumowanie selekcji krokowej							
Krok	Efekt Wstawione	Efekt Usunięte	DF	Liczba w	Chi-kwadrat punktacji	Chi-kwadrat Walda	Pr. > chi-kw.
1	trstprl		10	1	149.9788		<.0001
2	health		4	2	48.7410		<.0001
3	agea		1	3	39.9835		<.0001
4	njbspv		1	4	17.8241		<.0001
5	hinctnta		9	5	32.6361		0.0002
6	euftf		10	6	27.2039		0.0024
7	happy		10	7	21.8557		0.0159

Źródło: Opracowanie własne

Ostateczny model zawiera 7 zmiennych istotnych statystycznie. Zmienne STFDEM (ocena funkcjonowania demokracji), IPSUCES, EDUYRS okazały się nieistotna.

Tabela 9. Wyniki efektów typu 3

Analiza efektów typu 3			
Efekt	DF	Chi-kwadrat Walda	Pr. > chi-kw.
agea	1	43.6869	<.0001
njbspv	1	17.9819	<.0001
happy	10	20.9195	0.0217
euftf	10	26.9011	0.0027
health	4	38.4838	<.0001
hinctnta	9	29.7331	0.0005
trstprl	10	83.6300	<.0001

Źródło: Opracowanie własne

Wszystkie wartości p-value dla zmiennych przedstawionych w Tabeli 10 sugerują, że należy odrzucić hipotezę zerową na rzecz hipotezy alternatywnej, mówiącej o tym, że oceny parametrów istotnie różnią się od zera, co pozwala na ich uzasadnioną interpretację.

Na podstawie poniższych tabeli można dokonać interpretacji poszczególnych oszacowań parametrów. Ze względu na spełnienie założenia proporcjonalności szans interpretacja szans zarówno dla sytuacji bycia nieufnym w porównaniu do bycia średnioufnym bądź ufnym jest taka sama jak dla sytuacji bycia ufnym w porównaniu do bycia średnioufnym bądź nieufnym.

Tabela 10. Analiza ocen maksymalnej wiarygodności

Analiza ocen maksymalnej wiarygodności							
Parametr		DF	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.	Exp(Est)
Intercept	01: 1	1	0.1774	0.1965	0.8150	0.3666	1.194
Intercept	02: 2	1	2.0050	0.2018	98.7066	<.0001	7.426
agea		1	-0.0169	0.00256	43.6869	<.0001	0.983
njbspv		1	-0.0227	0.00534	17.9819	<.0001	0.978
happy	0	1	0.9459	0.7297	1.6805	0.1949	2.575
happy	1	1	1.5868	0.7154	4.9202	0.0265	4.888
happy	2	1	0.0292	0.4094	0.0051	0.9431	1.030
happy	3	1	0.0782	0.3521	0.0494	0.8242	1.081
happy	4	1	-0.3395	0.2711	1.5684	0.2104	0.712
happy	5	1	-0.2072	0.1961	1.1162	0.2907	0.813
happy	6	1	-0.1491	0.1921	0.6028	0.4375	0.861
happy	7	1	-0.4163	0.1559	7.1303	0.0076	0.660
happy	8	1	-0.5093	0.1463	12.1190	0.0005	0.601
happy	9	1	-0.6514	0.1589	16.8019	<.0001	0.521
euftf	0	1	0.4192	0.1369	9.3712	0.0022	1.521
euftf	1	1	0.2569	0.1729	2.2084	0.1373	1.293
euftf	2	1	0.0948	0.1409	0.4526	0.5011	1.099
euftf	3	1	0.3497	0.1286	7.3994	0.0065	1.419
euftf	4	1	0.1805	0.1360	1.7602	0.1846	1.198
euftf	5	1	0.0469	0.0945	0.2465	0.6196	1.048
euftf	6	1	-0.2438	0.1497	2.6534	0.1033	0.784
euftf	7	1	-0.1282	0.1624	0.6227	0.4300	0.880
euftf	8	1	-0.4124	0.1946	4.4928	0.0340	0.662
euftf	9	1	-0.1269	0.2783	0.2081	0.6482	0.881
health	1	1	-0.5138	0.1122	20.9883	<.0001	0.598
health	2	1	-0.4355	0.1001	18.9373	<.0001	0.647
health	3	1	0.1155	0.1073	1.1583	0.2818	1.122
health	4	1	0.3142	0.1581	3.9509	0.0468	1.369
hinctnta	1	1	0.1910	0.1209	2.4947	0.1142	1.210
hinctnta	2	1	0.3146	0.1225	6.5937	0.0102	1.370
hinctnta	3	1	0.1906	0.1170	2.6515	0.1035	1.210
hinctnta	4	1	0.2666	0.1417	3.5381	0.0600	1.306
hinctnta	5	1	0.1748	0.1338	1.7064	0.1915	1.191
hinctnta	6	1	-0.2215	0.1422	2.4245	0.1194	0.801
hinctnta	7	1	-0.1576	0.1380	1.3048	0.2533	0.854
hinctnta	8	1	-0.4147	0.1521	7.4318	0.0064	0.661
hinctnta	9	1	-0.4137	0.1472	7.8985	0.0049	0.661
trstprl	0	1	0.8307	0.1575	27.8149	<.0001	2.295
trstprl	1	1	0.5878	0.2235	6.9153	0.0085	1.800
trstprl	2	1	0.5570	0.1512	13.5723	0.0002	1.745
trstprl	3	1	0.2657	0.1350	3.8707	0.0491	1.304
trstprl	4	1	0.2031	0.1256	2.6148	0.1059	1.225
trstprl	5	1	0.0198	0.1103	0.0322	0.8575	1.020
trstprl	6	1	-0.2732	0.1306	4.3734	0.0365	0.761
trstprl	7	1	-0.1436	0.1304	1.2133	0.2707	0.866
trstprl	8	1	-0.8391	0.1634	26.3792	<.0001	0.432
trstprl	9	1	-0.7757	0.3623	4.5839	0.0323	0.460

Źródło: Opracowanie własne

Interpretacja - stała:

Gdyby stwierdzenie ,,wszystkie inne zmienne przyjmują wartość 0" miało sens, interpretacja stałej byłaby następująca: dla kategorii 1 szanse, że respondent będzie nieufny, a nie średnioufny i ufny są jak 12:10. Dla kategorii 2, szanse, że respondent będzie średnioufny, a nie ufny są jak 7,5:1.

Tabela 11. Analiza ocen maksymalnej wiarygodności przy dopasowaniu modelu do poziomu 3

Analiza ocen maksymalnej wiarygodności							
Parametr		DF	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.	Exp(Est)
Intercept	03: 3	1	-2.0050	0.2018	98.7066	<.0001	0.135
Intercept	02: 2	1	-0.1774	0.1965	0.8150	0.3666	0.837

Źródło: Opracowanie własne

Przy dopasowaniu modelu do kategorii 3 interpretacja wygląda następująco: dla kategorii 3 szansa, że respondent będzie średnioufny lub nie ufny w relacji do ufnego to 13:100. Dla kategorii 2 szasa, że respondent będzie nieufny, a nie średnioufny lub ufny to 83:100.

Interpretacja- zmienne ciągłe:

Jeśli osoby będą różniły się tylko pod względem wieku, to osoba starsza o rok będzie miała o 2% mniejsze szanse na bycie ufnym niż osoba młodsza.

Jeśli osoby będą się różnić pod względem liczby podwładnych, to osoba z 1 więcej podwładnym będzie miała o 2% mniejsze szanse na bycie ufnym.

Tabela 12. Ocena ilorazów szans

Oceny ilorazów szans i przedziały ufności Walda				
Efekt	Jednostka	Ocena	Przedział ufności 95%	
agea	1.0000	0.983	0.978	0.988
njbspv	1.0000	0.978	0.967	0.988
happy 0 od 10	1.0000	3.719	0.761	18.180
happy 1 od 10	1.0000	7.059	1.495	33.317
happy 2 od 10	1.0000	1.487	0.609	3.629
happy 3 od 10	1.0000	1.561	0.727	3.354
happy 4 od 10	1.0000	1.028	0.571	1.852
happy 5 od 10	1.0000	1.174	0.772	1.786
happy 6 od 10	1.0000	1.244	0.829	1.867
happy 7 od 10	1.0000	0.952	0.691	1.313
happy 8 od 10	1.0000	0.868	0.648	1.162
happy 9 od 10	1.0000	0.753	0.548	1.033
euftf 0 od 10	1.0000	2.353	1.308	4.234
euftf 1 od 10	1.0000	2.001	1.060	3.778
euftf 2 od 10	1.0000	1.701	0.936	3.093
euftf 3 od 10	1.0000	2.195	1.226	3.931
euftf 4 od 10	1.0000	1.854	1.027	3.346
euftf 5 od 10	1.0000	1.622	0.937	2.808
euftf 6 od 10	1.0000	1.213	0.664	2.215
euftf 7 od 10	1.0000	1.361	0.733	2.528
euftf 8 od 10	1.0000	1.025	0.529	1.984
euftf 9 od 10	1.0000	1.363	0.619	3.003
health 1 od 5	1.0000	0.356	0.179	0.708
health 2 od 5	1.0000	0.385	0.196	0.754
health 3 od 5	1.0000	0.668	0.341	1.308
health 4 od 5	1.0000	0.814	0.393	1.685
hinctnta 1 od 10	1.0000	1.129	0.767	1.661
hinctnta 2 od 10	1.0000	1.277	0.868	1.880
hinctnta 3 od 10	1.0000	1.128	0.773	1.647
hinctnta 4 od 10	1.0000	1.217	0.803	1.846
hinctnta 5 od 10	1.0000	1.111	0.745	1.656
hinctnta 6 od 10	1.0000	0.747	0.495	1.128
hinctnta 7 od 10	1.0000	0.796	0.531	1.196
hinctnta 8 od 10	1.0000	0.616	0.401	0.946
hinctnta 9 od 10	1.0000	0.617	0.406	0.937
trstprl 0 od 10	1.0000	3.537	1.570	7.968
trstprl 1 od 10	1.0000	2.774	1.144	6.726
trstprl 2 od 10	1.0000	2.690	1.197	6.045
trstprl 3 od 10	1.0000	2.010	0.908	4.449
trstprl 4 od 10	1.0000	1.888	0.858	4.156
trstprl 5 od 10	1.0000	1.572	0.723	3.418
trstprl 6 od 10	1.0000	1.173	0.530	2.594
trstprl 7 od 10	1.0000	1.335	0.605	2.946
trstprl 8 od 10	1.0000	0.666	0.294	1.510
trstprl 9 od 10	1.0000	0.709	0.241	2.086

Źródło: Opracowanie własne

Interpretacja - zmienne kategoryczna o kodowaniu typu efekt:

Kodowanie typu efekt nie umożliwia interpretacji wartości funkcji wykładniczej parametrów w kategoriach ilorazu szans. Wartość przydatną w zakresie interpretacji można znaleźć w tabeli powyżej. Żadna z kategorii nie jest istotna statystycznie.

3.2. Ocena jakości modelu regresji uporządkowanej

Do oceny modelu uporządkowanej regresji logistycznej wykorzystano:

- Współczynnik determinacji R2
- Procent zgodnych, niezgodnych obserwacji
- Statystyki D Somers, Gamma
- Statystka c, określająca pole pod krzywą ROC
- Kryterium Akaike, Bayesa-Schwartza oraz -2 log L

Współczynnik determinacji R^2 to jedna z podstawowych miar jakości dopasowania modelu. Informuje o tym, jaka część zmienności zmiennej objaśnianej została wyjaśniona przez model. Jest miarą stopnia, w jakim model wyjaśnia kształtowanie się zmiennej objaśnianej.

Tabela 13. Współczynnik R-kwadrat

R-kwadrat	0.1634	Maksymalnie przeskalowane R-kwadrat	0.1853

Źródło: Opracowanie własne

Współczynnik determinacji R^2 wyniósł 0,1634. Jakość dopasowania modelu do danych jest w omawianym modelu dość dobra (jak na model regresji logistycznej), Należy mieć na uwadze fakt, że niskie wartości współczynnika determinacji spowodowane są modelowaniem zjawiska na poziomie indywidualnym (mikroekonomicznym). Według literatury modele oparte na tego typu danych z reguły osiągają wartości z przedziału 5%-40%.

Tabela 14. Miary skojarzenia

Skojarzenie prognozowanych prawdopodobieństw i obserwowanych odpowiedzi			
Procent zgodnych	69.0	D Somersa	0.383
Procent niezgodnych	30.7	Gamma	0.385
Procent równych	0.3	Tau-a	0.248
Pary	1243118	c	0.692

Źródło: Opracowanie własne

Miary skojarzenia badające zdolności predykcyjne modelu wskazują na 69% poprawnie zaklasyfikowanych obserwacji, 30,7% źle zaklasyfikowanych i 0,3% przypadków bez decyzji. Statystyki D Somersa, Gamma oraz Tau-a na podstawie tablic kontyngencji testując niezależność zmiennych objaśnianej i objaśniających. Statystyka Gamma mówi o proporcjonalnej redukcji błędu, czyli nadwyżce zgodnych par ponad niezgodne wyrażana jako procent w stosunku do wszystkich wyodrębnionych par i wynosi 38%. Statystyka D

Sommera uwzględnia też pary związane, których liczba w omawianym modelu jest bardzo niska, dlatego wartość tej statystyki jest zbliżona do statystyki Gamma. Statystyka c określa pole pod krzywą ROC. Statystyka c powstała jako zależność pomiędzy zgodnymi skojarzeniami par a popełnionymi przez model niezgodnościami przyporządkowań. Im większa wartość c tym model lepiej przewiduje wartości zmiennej zależnej. Dla tego modelu wartość pola pod krzywą ROC wyniosła 0,69.

Kolejnym etapem była analiza kryteriów informacyjnych: Akaike, Bayesa-Schwartza oraz -2 log L. Stosując kryteria informacyjne do wyboru modelu spośród zbioru modeli - kandydatów wybiera się ten model, któremu odpowiada minimalna wartość danego kryterium informacyjnego.

Tabela 15. Statystyki dopasowania

Statystyki dopasowania		
Kryterium	**Tylko wyraz wolny**	**Wyraz wolny i współzmienne**
AIC	4191.556	3932.030
SC	4202.716	4194.299
-2 log L	4187.556	3838.030

Źródło: Opracowanie własne

Na podstawie kryteriów informacyjnych nie da się wybrać modelu z optymalną liczbą zmiennych, da się jedynie porównać jakość przynajmniej dwóch modeli. Warto przyznać, że wartości AIC i SC powinny być jak najniższe.

4. Model regresji binarnej typu forward

Następnym modelem regresji opracowanym w tym referacie jest model regresji binarnej typu forward. Jest to metoda, która polega na stopniowym dołączaniu do modelu kolejnych zmiennych. W pierwszym kroku tworzony jest model bez zmiennych przyczynowych. Zmienne są kolejno dodawane do modelu według malejącej wartości statystyki resztowej (i rosnącego p-value dla tej statystyki). W każdym kroku do modelu dodawana jest ta zmienna, której statystyka resztowa ma najmniejsze p-value. Zmienne są dodawane dopóty, dopóki p-value jest mniejsze od określonej wartości maksymalnej (kryterium wejścia). Algorytm kończy się, kiedy dodanie dowolnej z pozostałych zmiennych powoduje, że p-value jest większe niż kryterium wejścia.

Metoda ta ma dwa ograniczenia. Niektóre ze zmiennych nigdy nie trafiają do modelu, przez co ich istotność nie jest nigdy określona. Drugim ograniczeniem jest to, że zmienna, która została dodana do modelu, pozostaje w nim już do końca algorytmu, nawet jeśli utraciła istotność po dodaniu innych zmiennych.

Zmienna PPLFAIR została przekonwertowana na zmienną binarną z wartościami:

- 0 dla kategorii 1 i 2 (nieufny i średnioufny). Można tą wartość potraktować jako „raczej nieufny"
- 1 dla kategorii 3 (ufny).

Charakterystyka zbioru została przedstawiona w rozdziale 2. Dla prezentowanego modelu poziom istotności określono jako 0,05.

W przedstawionym modelu, każda ze zmiennych posiada tylko 1 stopień swobody. Autor uznał, że dla potrzeb łatwiejszej interpretacji wyników jeden z modeli przedstawionych w referacie powinien posiadać właśnie taki rodzaj kodowania zmiennych.

4.1 Wyniki modelu regresji logistycznej binarnej typu forward

Tabela 16. Podstawowe wyniki modelu forward

Wczytano obserwacji	1959
Użyto obserwacji	1959

Profil odpowiedzi		
Wartość uporządkowana	_RESPONSE	Całkowita liczebność
1	01: inne	1151
2	02: trojki	808

Źródło: Opracowanie własne

Charakterystyka zbioru została wspomniana w poprzednich rozdziałach. Różnicą jest tu wartość zmiennej zależnej: są tylko 2 wartości uporządkowanej. Liczebność pierwszej wynosi 1151, drugiej 808. Nie są to ilości bliźniacze, ale autor uznał je za wystarczająco zbliżone i optymalne przy tworzeniu analizy.

Status zbieżności- kryterium zbieżności został spełnione.

Tabela 17. Test Beta=0 dla forward

Testowanie globalnej hipotezy zerowej: BETA=0			
Test	Chi-kwadrat	DF	Pr. > chi-kw.
Iloraz wiarygod.	230.4723	9	<.0001
Wynik punktowy	214.9723	9	<.0001
Wald	192.9927	9	<.0001

Źródło: Opracowanie własne

Weryfikacja istotności parametrów modelu testując globalną hipotezę zerową. Wyniki testów, zarówno Walda jak i Score, przyjmują wartości p-value na poziomie niższym niż 0,001, co sugeruje odrzucenie hipotezy zerowej na rzecz alternatywnej.

Tabela 18. Wyniki selekcji postępującej (forward)

Podsumowanie selekcji postępującej					
Krok	Efekt wstawiony	DF	Liczba w	Chi-kwadrat punktacji	Pr. > chi-kw.
1	trstprl	1	2	99.7235	<.0001
2	health	1	3	46.3854	<.0001
3	hinctnta	1	4	21.3347	<.0001
4	eufttf	1	5	10.4709	0.0012
5	njbspv	1	6	12.8885	0.0003
6	happy	1	7	8.7839	0.0030
7	ipsuces	1	8	4.0388	0.0445
8	eduyrs	1	9	3.9016	0.0482

Źródło: Opracowanie własne

Model ostateczny zawiera 8 zmiennych istotnych statystycznie. Każda w tym modelu posiada tylko 1 stopień swobody, co ułatwia interpretację. Iteracja nr 9 zakończyła się na poziomie p-value 0.048, to jest poniżej ustalonego progu 0.05. Zmienne najbardziej statystycznie istotne to TRSTPRL, HEALTH i HINCTNTA gdzie p-value jest poniżej 0.001.

W poniższej tabeli uwzględniono analizę ocen maksymalnej wiarygodności. Po wynikach modelu można dojść do następujących wniosków:

- wyraz wolny dla tego modelu wynosi 3,5. Oznacz to, że gdyby wszystkie pozostałe zmienne miały wartość 0, taki byłby poziom zaufania respondenta
- respondent starszy o 1 rok, będzie miał o 2% mniejsze zaufanie, niż respondent rok młodszy
- respondent uczęszczający do szkoły o 1 rok dłużej, będzie miał o 3% mniejsze zaufanie, w porównaniu z respondentem, który uczęszczał o 1 rok krócej
- respondent, który wierzy w integrację europejską o 1/10 bardziej, będzie miał o 7% mniejsze zaufanie do ludzi

- respondent, który będzie miał o 1/5 gorsze zdrowie, będzie ufał ludziom o 28% bardziej
- respondent, który uważa że sukcesy są ważne o 1/6 bardziej, będzie ufał ludziom 8% bardziej
- respondent, który wierzy w parlament o 1/10 bardziej, będzie ufał ludziom o 15% mniej
- respondent, który ma o 1 podwładnego więcej, będzie ufał ludziom o 2% mniej
- respondent, który jest o 1/10 bardziej szczęśliwy, będzie ufał ludziom o 9% mniej

Tabela 19. Analiza ocen maksymalnej wiarygodności dla forward

Analiza ocen maksymalnej wiarygodności					
Parametr	DF	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.
Intercept	1	3.5082	0.4149	71.4935	<.0001
agea	1	-0.0163	0.00298	29.7890	<.0001
eduyrs	1	-0.0269	0.0137	3.8572	0.0495
euftf	1	-0.0672	0.0202	11.0849	0.0009
health	1	0.2755	0.0599	21.1501	<.0001
ipsuces	1	-0.0765	0.0375	4.1633	0.0413
hinctnta	1	-0.0610	0.0188	10.5717	0.0011
trstprl	1	-0.1581	0.0220	51.5526	<.0001
njbspv	1	-0.0203	0.00568	12.7402	0.0004
happy	1	-0.0948	0.0311	9.2752	0.0023

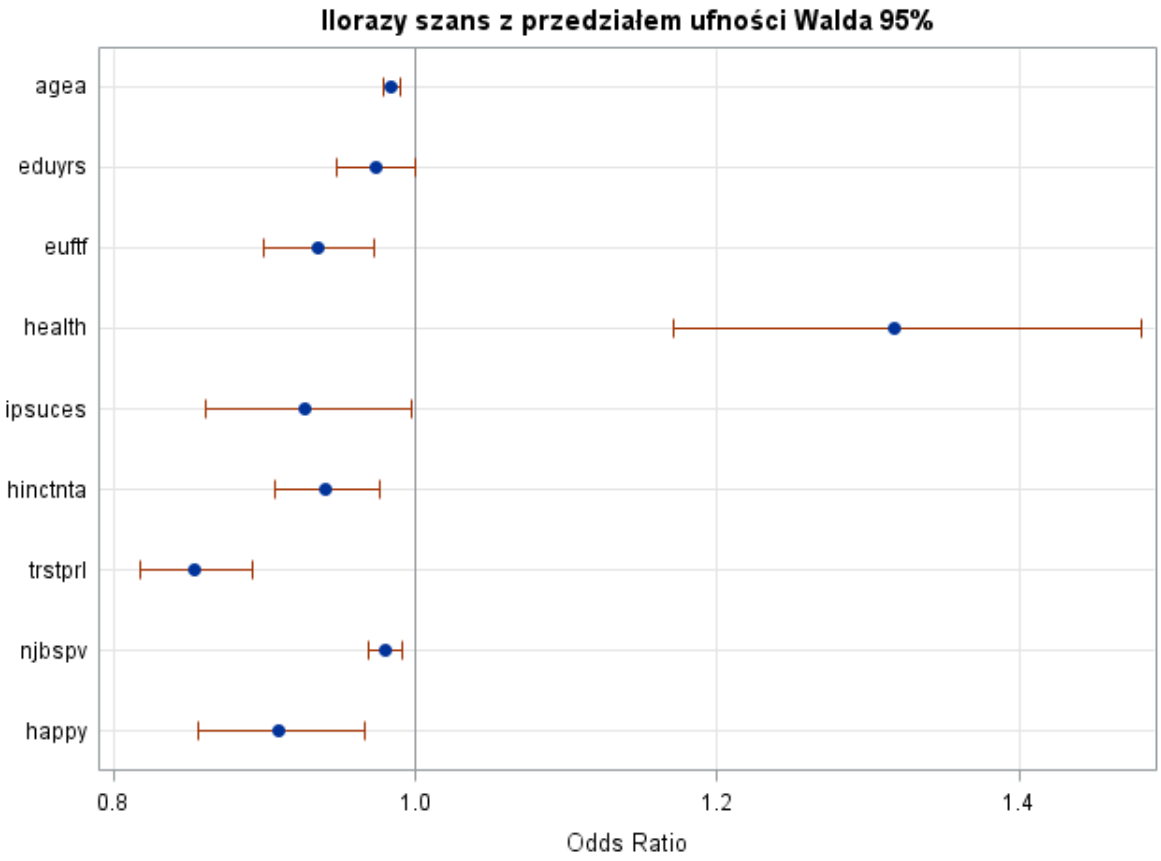
Źródło: Opracowanie własne

Tabela 20. Oceny ilorazu szans dla forward

Oceny ilorazu szans			
Efekt	Wynik punktowy	Przedział ufności Walda 95%	
agea	0.984	0.978	0.990
eduyrs	0.973	0.948	1.000
euftf	0.935	0.899	0.973
health	1.317	1.171	1.481
ipsuces	0.926	0.861	0.997
hinctnta	0.941	0.907	0.976
trstprl	0.854	0.818	0.891
njbspv	0.980	0.969	0.991
happy	0.910	0.856	0.967

Źródło: Opracowanie własne

Rysunek 13. Ilorazy szans



Źródło: Opracowanie własne

4.2. Ocena jakości modelu regresji binarnej typu forward

Do oceny jakości modelu użyto tych samych miar, co do regresji uporządkowanej.

Tabela 21. Współczynnik R-kwadrat dla forward

R-kwadrat	0.1110	Maksymalnie przeskalowane R-kwadrat	0.1495

Źródło: Opracowanie własne

Współczynnik determinacji R^2 wyniósł 0,1495. Jakość dopasowania modelu do danych jest w omawianym modelu dobra.

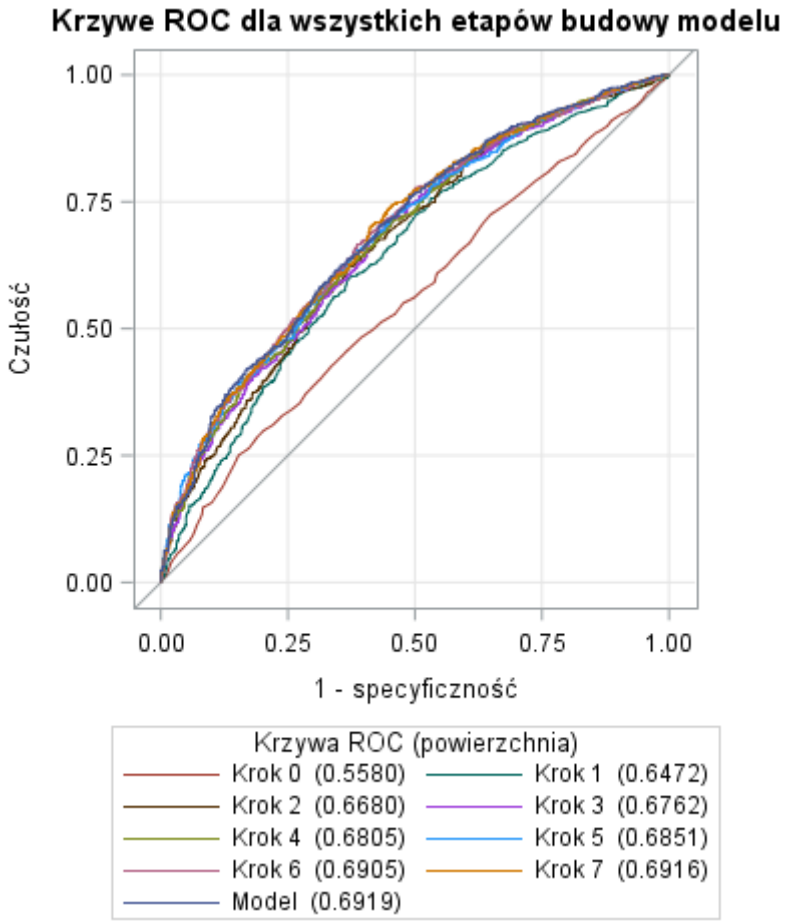
Tabela 22. Miary skojarzenia dla forward

Skojarzenie prognozowanych prawdopodobieństw i obserwowanych odpowiedzi			
Procent zgodnych	69.2	D Somersa	0.384
Procent niezgodnych	30.8	Gamma	0.384
Procent równych	0.0	Tau-a	0.186
Pary	930008	c	0.692

Źródło: Opracowanie własne

Miary skojarzenia badające zdolności predykcyjne modelu wskazują na 69,2% poprawnie zaklasyfikowanych obserwacji, 30,8% źle zaklasyfikowanych i 0,4% przypadków bez decyzji. Statystyki D Somersa, Gamma oraz Tau-a na podstawie tablic kontyngencji testując niezależność zmiennych objaśnianej i objaśniających. Statystyka Gamma mówi o proporcjonalnej redukcji błędu, czyli nadwyżce zgodnych par ponad niezgodne wyrażana jako procent w stosunku do wszystkich wyodrębnionych par i wynosi 38 %. Statystyka D Sommera uwzględnia też pary związane, których liczba w omawianym modelu jest bardzo niska, dlatego wartość tej statystyki jest zbliżona do statystyki Gamma. Statystyka c określa pole pod krzywą ROC. Statystyka c powstała jako zależność pomiędzy zgodnymi skojarzeniami par a popełnionymi przez model niezgodnościami przyporządkowań. Im większa wartość c tym model lepiej przewiduje wartości zmiennej zależnej. Dla modelu forward po 8 iteracjach wartość pola pod krzywą ROC wyniosła 0,692.

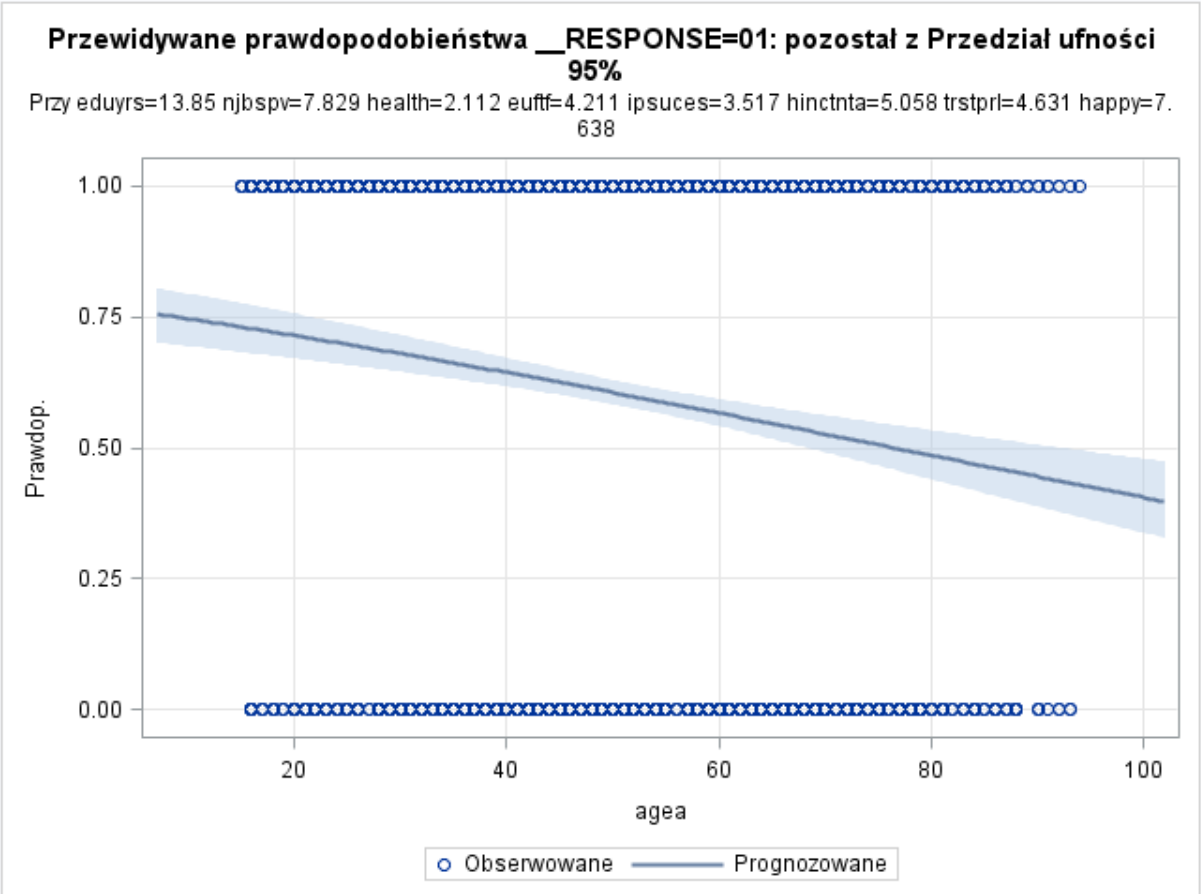
Rysunek 14. Krzywe ROC dla każdej z iteracji



Źródło: Opracowanie własne

Wykres krzywej ROC ukazuje jak jakość dopasowania modelu poprawiała się wraz z każdą iteracją. Można również zauważyć, ze już od kroku 2. do ostatniego jakość modelu była zbliżona.

Rysunek 15. Przewidywane prawdopodobieństwa



Źródło: Opracowanie własne

Na wykresie można zobaczyć, jak funkcja prawdopodobieństwa maleje wraz z wiekiem. Oznacza to, że w wraz z wiekiem, respondenci stają się mniej ufni.

Tabela 23. Statystyki dopasowania

Statystyki dopasowania		
Kryterium	Tylko wyraz wolny	Wyraz wolny i współzmienne
AIC	2657.384	2444.912
SC	2662.965	2500.714
-2 log L	2655.384	2424.912

Źródło: Opracowanie własne

4.3 Ocena jakości modelu regresji binarnej typu backward

Dla potrzeb referatu autor przeprowadził również analizę z użyciem modelu regresji logistycznej typu backward. Metoda backward oznacza, że usuwane są najmniej istotne zmienne z modelu zawierającego wszystkie zmienne objaśniające dopóki wszystkie zmienne w modelu będą istotne.

Wyniki modelu okazały się bliźniaczo podobne jak w regresji typu forward. Jedyną różnicą jest liczba kroków- w metodzie backward przeprowadzono tylko jedna iterację- usunięcie zmiennej STFDEM.

5. Bibliografia

1. Allison P.D., “Logistic Regression Using SAS Theory and Application”, SAS Press

2. „Zaawansowane metody analiz statystycznych”, red. Naukowa Ewa Frątczak, Oficyna Wydawnicza SGH, Warszawa 2013

3. „Ekonometria i badania operacyjne”, red. naukowa m. Gruszczyński, T. Kuszewski, M. Podgórska, Wydawnictwo Naukowe PWN, Warszawa 2013

6. Spis tabel i rysunków

7. Kody SAS

Regresja uporządkowana

```
/* ---------------------------------------------------------------
--
   Kod wygenerowany przez zadanie SAS-a

   Wygenerowany dnia: niedziela, 28 stycznia 2018 o godz. 18:33:10
   Przez zadanie: Regresja logistyczna

   Dane wejściowe: Local:WORK.RDATA
   Serwer:  Local
   ---------------------------------------------------------------
-- */
ODS GRAPHICS ON;

%_eg_conditional_dropds(WORK.SORTTempTableSorted,
            WORK.TMPMod);
/* ---------------------------------------------------------------
--
   Sortowanie zbioru Local:WORK.RDATA
   ---------------------------------------------------------------
-- */

PROC SQL;
      CREATE VIEW WORK.SORTTempTableSorted AS
```

```
            SELECT T.pplfair, T.agea, T.eduyrs, T.stfdem, T.njbspv,
T.happy, T.hinctnta, T.euftf, T.trstprl, T.ipsuces, T.health
      FROM WORK.RDATA as T
;
QUIT;

DATA WORK.TMPMod;
      SET WORK.SORTTempTableSorted;
      length __RESPONSE $ 10;
      IF pplfair=1 THEN __RESPONSE="01: 1";
      IF pplfair=2 THEN __RESPONSE="02: 2";
      IF pplfair=3 THEN __RESPONSE="03: 3";
RUN;

TITLE;
TITLE1 "Rezultaty regresji logistycznej";
FOOTNOTE;
FOOTNOTE1 "Wygenerowane przez System SAS (&_SASSERVERNAME, &SYSSCPL)
dnia %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) o godz.
%TRIM(%SYSFUNC(TIME(), NLTIMAP20.))";
PROC LOGISTIC DATA=WORK.TMPMod
            PLOTS(ONLY)=ALL
      ;
      CLASS happy      (PARAM=EFFECT) hinctnta     (PARAM=EFFECT)
euftf       (PARAM=EFFECT) trstprl      (PARAM=EFFECT) ipsuces
      (PARAM=EFFECT) health       (PARAM=EFFECT);
      MODEL __RESPONSE (DESCENDING)=agea eduyrs stfdem njbspv happy
euftf health hinctnta trstprl ipsuces       / expb
            SELECTION=STEPWISE
            SLE=0.05
            SLS=0.05
            INCLUDE=0
            RSQUARE
            LINK=LOGIT
            CLPARM=WALD
            CLODDS=WALD
            ALPHA=0.05
      ;
RUN;
QUIT;

/* ----------------------------------------------------------------
--
   Koniec kodu zadania
   ----------------------------------------------------------------
-- */
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORTTempTableSorted,
            WORK.TMPMod);
TITLE; FOOTNOTE;
ODS GRAPHICS OFF;
```

Regresja binarna typu forward:

```
ODS GRAPHICS ON;

%_eg_conditional_dropds(WORK.SORTTempTableSorted,
                WORK.TMPMod);

PROC SQL;
        CREATE VIEW WORK.SORTTempTableSorted AS
                SELECT T.pplfair, T.agea, T.eduyrs, T.euftf, T.happy, T.health, T.ipsuces, T.stfdem,
T.hinctnta, T.trstprl, T.njbspv
        FROM WORK.RDATA as T
;
QUIT;

DATA WORK.TMPMod;
        SET WORK.SORTTempTableSorted;
        length __RESPONSE $ 20;
        IF strip(put(pplfair,MOJA.))="inne" THEN __RESPONSE="01: inne";
        IF strip(put(pplfair,MOJA.))="trojki" THEN __RESPONSE="02: trojki";
RUN;

TITLE;
TITLE1 "Rezultaty regresji logistycznej";
FOOTNOTE;
FOOTNOTE1 "Wygenerowane przez System SAS (&_SASSERVERNAME, &SYSSCPL) dnia
%TRIM(%QSYSFUNC(DATE(), NLDATE20.)) o godz. %TRIM(%SYSFUNC(TIME(),
NLTIMAP20.))";
PROC LOGISTIC DATA=WORK.TMPMod
                PLOTS(ONLY)=ALL
        ;
        MODEL __RESPONSE=agea eduyrs euftf health ipsuces stfdem hinctnta trstprl njbspv happy
                /
                SELECTION=FORWARD
                SLE=0.05
                INCLUDE=1
                RSQUARE
                LINK=LOGIT
                CLPARM=WALD
                ALPHA=0.05
        ;
        FORMAT pplfair MOJA.;
RUN;
QUIT;

/* ---------------------------------------------------------------
  Koniec kodu zadania
  --------------------------------------------------------------- */
```

```
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORTTempTableSorted,
                  WORK.TMPMod);
TITLE; FOOTNOTE;
ODS GRAPHICS OFF;
```

Tworzenie formatu dla zmiennej PPLFAIR:

```
/* ---------------------------------------------------------------
   Kod wygenerowany przez zadanie SAS-a

   Wygenerowany dnia: czwartek, 25 stycznia 2018 o godz. 01:12:29
   Przez zadanie: Tworzenie formatu (moja - Local)

   Dane wejściowe: ."n
   Serwer:
   --------------------------------------------------------------- */

%_eg_conditional_dropds(WORK.CFMT_0000);

PROC FORMAT
        LIB=WORK
;
                VALUE moja ( DEFAULT=8 )
                        3 = "trzy"
                        OTHER = "pozostałe";
RUN;

RUN; QUIT;
TITLE; FOOTNOTE;
```

Inputowanie zbioru danych:

```
DATA  rdata ;
/* ---------------------------------------------------------------
--
    Proszę zmienić ścieżkę pliku na właściwą
    ---------------------------------------------------------------
-- */

INFILE  "C:/Users/Sebastian/Desktop/Analiza danych- Big data 2015-18/4. Semestr/Regresja
logistyczna/ECLIB000/mydata.txt"
    DSD
    LRECL= 31 ;
INPUT
 pplfair
 agea
 eduyrs
 euftf
 happy
```

```
 health
 ipsuces
 stfdem
 hinctnta
 trstprl
 njbspv
;
RUN;


proc print data= rdata;
run;
```

Korelacje:

```
/* ---------------------------------------------------------------------
--
   Kod wygenerowany przez zadanie SAS-a

   Wygenerowany dnia: niedziela, 28 stycznia 2018 o godz. 18:41:13
   Przez zadanie: Korelacje

   Dane wejściowe: Local:WORK.RDATA
   Serwer:  Local
   ---------------------------------------------------------------------
-- */
ODS GRAPHICS ON;

%_eg_conditional_dropds(WORK.SORTTempTableSorted);
/* ---------------------------------------------------------------------
--
   Sortowanie zbioru Local:WORK.RDATA
   ---------------------------------------------------------------------
-- */

PROC SQL;
      CREATE VIEW WORK.SORTTempTableSorted AS
            SELECT T.pplfair, T.agea, T.eduyrs, T.euftf, T.happy,
T.health, T.ipsuces, T.stfdem, T.hinctnta, T.trstprl, T.njbspv
      FROM WORK.RDATA as T
;
QUIT;
TITLE;
TITLE1 "Analiza korelacji";
FOOTNOTE;
FOOTNOTE1 "Wygenerowane przez System SAS (&_SASSERVERNAME, &SYSSCPL)
dnia %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) o godz.
%TRIM(%SYSFUNC(TIME(), NLTIMAP20.))";
PROC CORR DATA=WORK.SORTTempTableSorted
      PLOTS=NONE
      COV
```

```
      PEARSON
      KENDALL
      VARDEF=DF
      ;
      VAR pplfair agea eduyrs euftf happy health ipsuces stfdem
hinctnta trstprl njbspv;
RUN;

/* ---------------------------------------------------------------
--
   Koniec kodu zadania
   ---------------------------------------------------------------
-- */
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORTTempTableSorted);
TITLE; FOOTNOTE;
ODS GRAPHICS OFF;
```