WOJSKOWA AKADEMIA TECHNICZNA

im. Jarosława Dąbrowskiego

WYDZIAŁ CYBERNETYKI

SPRAWOZDANIE
Metody Eksploracji Danych

Temat laboratorium: **MODELE LOGITOWE. REGRESJA LOGISTYCZNA**

INFORMATYKA

(kierunek studiów)

INŻYNIERIA SYSTEMÓW – ANALIZA DANYCH

(specjalność)

Zespół:

Michał ŚLĘZAK
Szymon OLEŚKIEWICZ

Prowadzący laboratorium:

Dr inż. Romuald Hoffmann, prof. WAT

Warszawa 2025

Spis treści

Rozdział I. Zadanie 3 – Płatki śniadaniowe

I.1. Wykorzystane narzędzia i zależności

I.1.1. Modele

W celu wyznaczenia modeli logitowych wraz z parametrami na podstawie wykładów oraz wiedzy własnej przygotowano program w języku Python wykorzystujący biblioteki Pandas, NumPy, stastmodel, scikit-learn oraz scipy, które również zostały wykorzystane do obliczeń oraz wizualizacji wyników. Modele oraz ich parametry zostały wyznaczone za pomocą wyżej wspomnianych narzędzi

Warto jednak zaznaczyć, że w tym laboratorium, w celu zbadania zależności zmiennej dychotomicznej (w tym zadaniu zmienną tą będzie przynależność danych płatków do danej półki lub też nie) od zmiennych objaśniających (skład płatków).

Zastosowano model logitowy. Przewiduje on logarytm szans. W naszym przypadku, prawdopodobieństwo wystąpienia danych płatków na danej półce oznaczono jako p, a $X_1, X_2, \ldots, X_k$ są zmiennymi objaśniającymi (w naszym przypadku są to składniki płatków) natomiast $b_0, b_1, \ldots, b_k$ są to współczynniki przy tych zmiennych.

$$\hat{L} = \ln \frac{\hat{p}}{1-\hat{p}} = b_0 + \sum_{i=1}^{k} b_i \cdot X_i$$

Gdzie p to prawdopodobieństwo sukcesu (bycia na danej półce), iloraz w logarytmie naturalnym to iloraz szans (stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki) a L to nasza liniowa postać modelu logitowego po oszacowaniu. Parametr b_i (i od 1 do k) mówi nam o zmianie logarytmu szans przy wzroście zawartości danego składnika odżywczego o jednostkę. Warto zaznaczyć, że stosując przekształcenie odwrotne do transformacji powyższego modelu, otrzymamy oszacowanie prawdopodobieństwa p w postaci funkcji logistycznej.

$$\hat{p} = \frac{1}{1+e^{-\hat{L}}} = \frac{1}{1+e^{-(b_0+\sum_{i=1}^{k} b_i \cdot X_i)}}$$

I.1.2. Ocena istotności zmiennych – test Walda

Do oceny istotności statystycznej poszczególnych zmiennych w modelu wykorzystano test Walda. Statystyka z sprawdza hipotezę zerową $H_0: b_i = 0$. Jeśli wartość wartość (p-value, w naszym programie oznaczane jako $P > |z|$) jest mniejsza od przyjętego poziomu istotności (w naszym wypadku $\alpha = 0{,}05$), zmienną uznaje się za istotną statystycznie. Wartości krytyczne statystyk obliczono z

wykorzystaniem biblioteki Python – scipy. Statystyka Walda W_i jest równa kwadratowi z. Warto dodać, że statystyka Walda ma w przybliżeniu rozkład chi-kwadrat z liczbą stopni swobody równą 1 – to właśnie z tablic rozkładu chi-kwadrat będziemy brać wartość krytyczną.

I.1.3. Metryki oceny jakości klasyfikacji modeli

Do weryfikacji skuteczności modeli w procesie klasyfikacji przynależności danych płatków do danej półki wykorzystano wiele metryk, których implementacja znajduje się w bibliotekach Pythona. Metryki wykorzystane w obliczeniach znajdują się poniżej.

Macierz pomyłek, która jest zestawieniem wartości rzeczywistych naszego zbioru danych z wartościami przewidywanymi przez nasz model. Ma postać tabeli kwadratowej, w której wyróżnia się cztery główne wartości.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

True Positive – prawdziwie dodatnie: liczba przypadków, w których model poprawnie przewidział sukces (dane płatki są na danej półce i model to potwierdził)

True Negative – prawdziwie ujemne: liczba przypadków, w których model poprawnie przewidział brak zdarzenia (dane płatki nie są na danej półce i model to potwierdził).

False Positive – fałszywie dodatnie: model przewidział sukces, mimo że w rzeczywistości zdarzenie nie miało miejsca (dane płatki w rzeczywistości nie są na danej półce a model przewidział, że są).

False Negative – fałszywie ujemne: model przewidział brak zdarzenia, mimo że w rzeczywistości ono wystąpiło (dane płatki w rzeczywistości są na danej półce a model przewidział, że nie są).

Na podstawie tych 3 wartości są obliczane kolejne kluczowe metryki służące do oceny jakości modelu.

Dokładność (Accuracy) ukazuje nam ogólny odsetek poprawnych klasyfikacji (ile ogółem sklasyfikowano poprawnie).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Czułość (Recall) ukazuje nam zdolność modelu do wykrywania klasy pozytywnej – ile trafiłem z rzeczywistych płatków należących do danej półki (dane płatki należą do danej półki).

$$Recall = \frac{TP}{TP + FN}$$

Precyzja (Precision) ukazuje nam wiarygodność klasyfikacji pozytywnych – ile trafiłem naprawdę płatków należących do danej półki.

$$Precision = \frac{TP}{TP + FP}$$

Specyficzność (Specifity) ukazuje nam zdolność do wykrywania klasy negatywnej- ile trafiłem z rzeczywistych płatków nienależących do danej półki (płatki nie należą do danej półki).

$$Specifity = \frac{TN}{TN + FP}$$

F1-Score jest to średnia harmoniczna precyzji i czułości, która przydaje nam się w szczególności w niezbalansowanych klasach.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Dodatkowo, wykorzystano analizę krzywej ROC dla eksperymentu i weryfikacji poprawności klasyfikacji najlepszych wybranych modeli dla danej półki po wyłączeniu z danych treningowych producenta z o największej liczbie płatków w zbiorze danych (Receiver Operating Characrteristic). Jest ona graficzną reprezentacją skuteczności modelu klasyfikującego dla wszystkich możliwych progów klasyfikacji. Przedstawia ona zależność między 2 parametrami – czułością (True Positive Rate) oraz swoistością (False Positive Rate). Wykorzystano także

metrykę AUC, która jest polem pod krzywą ROC i która mówi nam o jakości modelu przed wyborem konkretnego progu klasyfikacji. Wykorzystując właśnie krzywą ROC można wyznaczyć optymalny próg klasyfikacji poprzez maksymalizację odległości (różnicy) TPR i FPR.

Jak już wcześniej wspomniano, do obliczeń wykorzystano narzędzia w postaci języka programowania Python, który, m.in. oblicza niektóre metryki.

Odchylenie standardowe modelu S_e uzyskamy pierwiastkując wariancję S_e^2.

$$S_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - k - 1}$$

Na podstawie wartości odchylenia standardowego S_e modelu jednej zmiennej można wyznaczyć odchylenia parametrów $a_0\ i\ a_1$, tak jak poniżej.

$$S_{a_0} = S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}$$

$$S_{a_1} = \frac{S_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}$$

Współczynnik Pseudo R^2 (McFaddena) pokazuje nam „dopasowanie modelu do danych". W regresji logistycznej jest to miara opierająca się na funkcji wiarygodności (Likelihood), która informuje nas o tym, o ile lepszy jest nasz model od modelu zerowego, czyli takiego zawierającego tylko wyraz wolny.

$$Pseudo\ R^2 = 1 - \frac{\ln L_M}{\ln L_0}$$

Gdzie w liczniku mamy logarytm funkcji wiarygodności dla modelu ze zmiennymi objaśniającymi a w mianowniku logarytm funkcji wiarygodności dla modelu tylko z wyrazem wolnym. Wzór na funkcję wiarygodności jest taki jak poniżej.

$$L(\theta|y, X) = \prod_i P(y_i|X_i, \theta)^{y_i} \cdot P(y_i = 1|X_i, \theta)^{1-y_i}$$

Gdzie:

$$\theta - wektor\ wartości\ parametrów\ modelu$$

$$y - wektor\ wartości\ zmiennej\ objaśnianej\ przyjmującej\ wartości\ 0\ lub\ 1$$

$$X - macierz\ zmiennych\ objaśniających$$

$$i - numer\ obserwacji$$

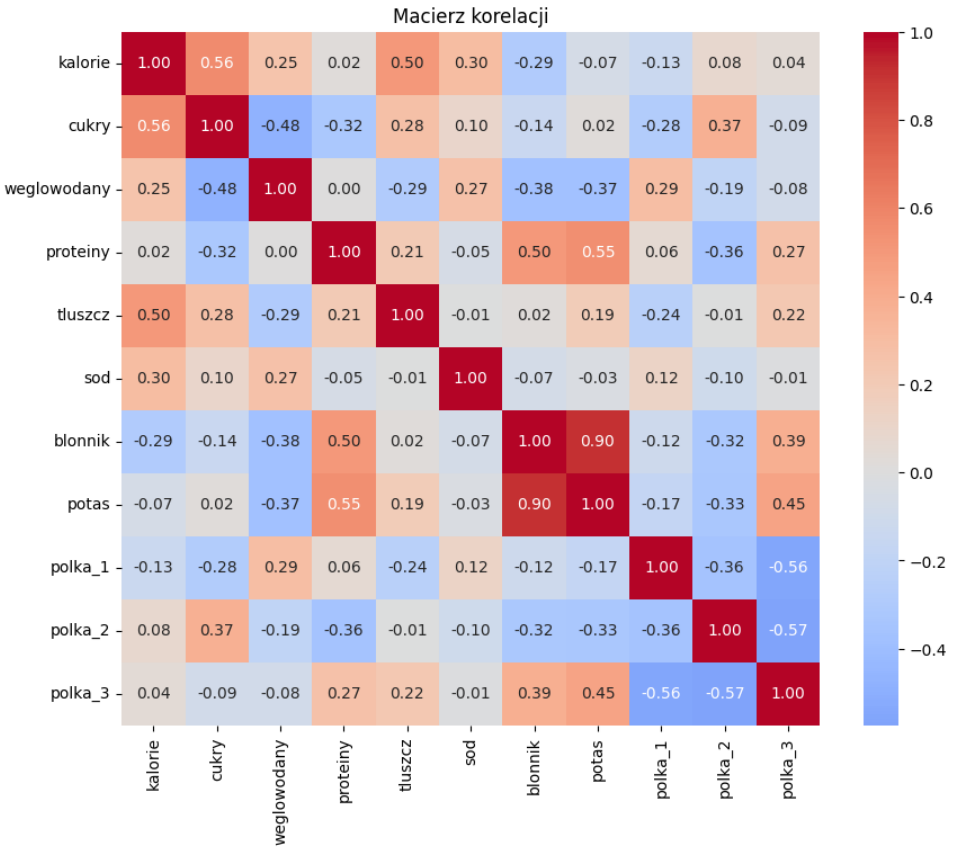
$P(y_i = 1|X_i, \theta) - prawdopodobieństwo, że\ dla\ i$
$- tej\ obserwacji\ y\ przyjmuje\ wartość\ 1, przy\ założeniu\ wartości\ zmiennej\ niezależnej\ X_i\ oraz$
$wartości\ parametrów\ regresji\ logistycznej.$

W procesie budowania modeli, dzielono dane na zbiór testowy i treningowy, gdyż liczność zbioru na to uzasadniała.

I.2. Eksploracja danych

I.2.1. Analiza korelacji między zmiennymi

W celu sprawdzenia korelacji między zmiennymi, wykorzystano macierz korelacji opartą o metodę Pearsona oraz stworzono jej wizualizację na mapie cieplnej.



Rys. 1. Wizualizacja macierzy korelacji między zmiennymi

Z analizy macierzy korelacji wynika, że półka 2 ma największą korelację ujemną z potasem, błonnikiem i białkiem, natomiast wysoką korelację dodatnią z cukrem, co może sugerować, że na półce 2 są płatki o najgorszym składzie, niezdrowe, których sklep chce się szybko pozbyć. Półka 3 ma największą korelację dodatnią z potasem, błonnikiem i białkiem natomiast wysoką korelację ujemną z cukrem, co może sugerować, że na półce 3 będą płatki o najlepszym składzie, tj. największej zawartości błonnika, potasu czy białka. Półka 1 nie ma jednoznacznej korelacji ze

składem, ma największą korelację ujemną z cukrem czy tłuszczem, ale tak samo ma korelację ujemną z błonnikiem czy potasem natomiast dodatnią korelację z białkiem (największa jest z węglowodanami). Można przypuszczać, że na tej półce będą płatki z wysoką zawartością węglowodanów, ale niską zawartością cukrów, gdyż te dwie zmienne mają skrajne, największe przeciwne korelacje z półką 1. Widoczna jest też dość wysoka korelacja między zmiennymi potas-błonnik-białko oraz cukier-kalorie, co może świadczyć o multikolinearności. Mając jednak na uwadze wartość biznesową, ludzie przeważnie patrzą w składzie płatków na cukier, kalorie, białko, błonnik i potas.

I.3. Modelowanie

W procesie analizy danych zdecydowano się postawić następujące pytania:

1. Czy skład płatków ma wpływ na ich umiejscowienie na półce w sklepie (1, 2 czy 3 półka)?
2. Czy płatki lepsze, zdrowsze, tj. np. z wyższą zawartością białka, potasu czy błonnika są umiejscowione wyżej?
3. Czy na półce środkowej (półka 2) są płatki gorsze jakościowo, których sklep chce się jak najszybciej pozbyć?
4. Jakie płatki są umieszczane na najniższej półce (półka 1)?

I.4. Modele dla półki 2

Półka ta znajduje się, zazwyczaj, na wysokości wzroku dzieci i w zasięgi ręku dorosłego szukającego szybkiego, niewymagającego wysiłku, wyboru (niekoniecznie najzdrowszego) płatków. Sugerując się macierzą korelacji, na tej półce najprawdopodobniej są płatki gorszej jakości, które sklep chce nam szybko sprzedać. Dlatego zdecydowano się na zbudowanie kilku modeli dla półki 2, aby sprawdzić te zależności. Przy wyborze zmiennych objaśniających kierowano się także macierzą korelacji.

I.4.1. Model logitowy – cukier jako zmienna objaśniająca

Modelem logitowym, który zbudowano jako pierwszy jest model oparty o zmienną objaśniającą cukier. Zdecydowano się na budowę tego modelu, aby sprawdzić czy wysoka zawartość cukru, jako niezdrowego składnika, jest jedynym czynnikiem, który sprawia, że dane płatki trafiają na półkę 2, co potwierdzałoby chociażby, np.

przyciąganie dzieci. Prawdopodobieństwo przynależności płatków do półki 2 oznaczono jako p a zmienną objaśniającą (zawartość cukru) jako X_1.

$$\hat{L} = \ln \frac{\hat{p}}{1-\hat{p}} = b_1 X_1 + b_0$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	p-value < 0,05	$W_i > \chi^2_{kryt}$
$b_0 = -2{,}9198$	0,762	0,000 – istotny	14,684 – istotny
$b_1 = 0{,}2576$	0,082	0,002 – istotny	9,910 – istotny

Tab. 1. Tabela z parametrami dla modelu logitowego 1

$$Pseudo\ R^2 = 17{,}4\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 81{,}3\%$$
$$Precision \approx 60\%$$
$$Recall \approx 75\%$$
$$Specifity \approx 83{,}3\%$$
$$F1 \approx 66{,}7\%$$
$$AUC \approx 0{,}67$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 3	FP = 2
Przewidziana 0	FN = 1	TN = 10

Tab. 2. Macierz pomyłek modelu logitowego 1

I.4.2. Model logitowy – cukier i kalorie jako zmienne objaśniające

Modelem logitowym, który zbudowano następnie jest model oparty o zmienne objaśniające cukier i kalorie. Zdecydowano się na budowę tego modelu, aby sprawdzić czy płatki wysokokaloryczne i słodkie są celowo umieszczane na tej półce,

aby np. pod wpływem impulsu, ludzie kupowali je jako szybki wybór na śniadanie, jako zastrzyk energii. Prawdopodobieństwo przynależności płatków do półki 2 oznaczono jako p a zmienne objaśniające (zawartość cukru i kalorii odpowiednio) jako X_1, X_2.

$$\hat{L} = \ln\frac{\hat{p}}{1-\hat{p}} = b_2X_2 + b_1X_1 + b_0$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -0{,}5907$	2,006	0,086 – nieistotny
$b_1 = 0{,}3183$	0,101	9,992 – istotny
$b_2 = -0{,}0260$	0,022	1,454 – nieistotny

Tab. 3. Tabela z parametrami dla modelu logitowego 2

$$Pseudo\ R^2 = 19{,}4\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 75\%$$
$$Precision \approx 50\%$$
$$Recall \approx 75\%$$
$$Specifity \approx 75\%$$
$$F1 \approx 60\%$$
$$AUC \approx 0{,}67$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 3	FP = 3
Przewidziana 0	FN = 1	TN = 9

Tab. 4. Macierz pomyłek modelu logitowego 2

I.4.3. Model logitowy – cukier, kalorie, potas, białko i błonnik jako zmienne objaśniające

Modelem logitowym, który zbudowano następnie jest model oparty o zmienne objaśniające cukier, kalorie, potas, białko i błonnik. Zdecydowano się na budowę tego modelu, aby sprawdzić czy na tej konkretnej półce liczy się tylko ten najgorszy skład płatków czy może obecność zdrowszych składników też ma na to wpływ co sklep chce nam szybko sprzedać poprzez umieszczenie płatków na półce 2. Prawdopodobieństwo przynależności płatków do półki 2 oznaczono jako p a zmienne objaśniające (zawartość cukru, kalorii, potasu, białka i błonnika odpowiednio) jako X_1, X_2, X_3, X_4, X_5.

$$\hat{L} = \ln \frac{\hat{p}}{1-\hat{p}} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = 0{,}1994$	2,767	0,005 – nieistotny
$b_1 = 0{,}2894$	0,114	6,396 – istotny
$b_2 = -0{,}0193$	0,030	0,415 – nieistotny
$b_3 = -0{,}0109$	0,015	0,524 – nieistotny
$b_4 = 0{,}0900$	0,455	0,039 – nieistotny
$b_5 = -0{,}4613$	0,539	0,731 – nieistotny

Tab. 5. Tabela z parametrami dla modelu logitowego 3

$$Pseudo\ R^2 = 31{,}7\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 81{,}3\%$$
$$Precision \approx 66{,}7\%$$
$$Recall \approx 50\%$$
$$Specifity \approx 91{,}7\%$$
$$F1 \approx 57{,}1\%$$

$$AUC \approx 0{,}67$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 2	FP = 1
Przewidziana 0	FN = 2	TN = 11

Tab. 6. Macierz pomyłek modelu logitowego 3

I.4.4. Model logitowy – cukier i białko jako zmienne objaśniające

Modelem logitowym, który zbudowano następnie jest model oparty o zmienne objaśniające cukier i białko. Zdecydowano się na budowę tego modelu, aby sprawdzić czy na tej konkretnej półce liczy się tylko ten najgorszy skład płatków czy może obecność białka jako składnika budulcowego mięśni oraz to, że większość osób patrzy na ten składnik też ma na to wpływ co sklep chce nam szybko sprzedać poprzez umieszczenie płatków na półce 2. Sugerując się macierzą korelacji, sprawdzamy czy większa zawartość cukru i mniejsza zawartość białka ma wpływ na umiejscowienie płatków na tej półce. Prawdopodobieństwo przynależności płatków do półki 2 oznaczono jako p a zmienne objaśniające (zawartość cukru i białka odpowiednio) jako X_1, X_2.

$$\hat{L} = \ln\frac{\hat{p}}{1-\hat{p}} = b_0 + b_1X_1 + b_2X_2$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -0{,}7388$	1,245	0,353 – nieistotny

$b_1 = 0{,}2101$	0,083	6,360 – istotny
$b_2 = -0{,}8$	0,415	3,717 – nieistotny

Tab. 7. Tabela z parametrami dla modelu logitowego 4

$$Pseudo\ R^2 = 23{,}4\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 81{,}3\%$$
$$Precision \approx 66{,}7\%$$
$$Recall \approx 50\%$$
$$Specifity \approx 91{,}7\%$$
$$F1 \approx 57{,}1\%$$
$$AUC \approx 0{,}70$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 2	FP = 1
Przewidziana 0	FN = 2	TN = 11

Tab. 8. Macierz pomyłek modelu logitowego 4

Na podstawie analizy wyników zbudowanych modeli dla półki 2, można stwierdzić, że żaden nie jest idealny. Jedynym modelem, którego metryki oceny są zadowalające i wysokie to model oparty tylko o zmienną objaśniającą – zawartość cukru w płatkach, co sugerowała także macierz korelacji (dodatnia korelacja). Dowodzi to także to, że cukier jest jedyną zmienną istotną statystycznie we wszystkich zbudowanych modelach – na podstawie testu Walda. Model ten charakteryzuje się wysoką czułością (~ 75%) oraz wysokim AUC (~ 0,67) a także najwyższym wskaźnikiem F1 (~ 66,7%). Analiza wyników pozwala stwierdzić, że prawdą jest jakoby na półce 2 znajdowały się płatki niezdrowe, o najgorszej jakości, o wysokiej zawartości cukru, których sklep chce się szybko „pozbyć" oraz, że docelowa grupa klientów dla płatków na tej półce to dzieci lub osoby, którym zależy na szybkim, niewymagającym wysiłku, wyborze płatków, niekoniecznie tym zdrowym.

I.5. Modele dla półki 3

Górna półka wymaga od klienta, zazwyczaj, wysiłku, bo nie znajduje się w zasięgu wzroku czy ręki dorosłego. Dlatego mając na uwadze macierz korelacji, najprawdopodobniej na tej półce są płatki najzdrowsze, tj. z wysoka zawartością białka, potasu czy błonnika. Dlatego zdecydowano się na zbudowanie kilku modeli dla półki 3, aby sprawdzić te zależności. Przy wyborze zmiennych objaśniających kierowano się także macierzą korelacji.

I.5.1. Model logitowy – potas, białko i błonnik jako zmienne objaśniające

Modelem logitowym, który zbudowano jako pierwszy jest model oparty o zmienne objaśniające potas, białko i błonnik. Zdecydowano się na budowę tego modelu, aby sprawdzić czy wysoka zawartość mieszanki zdrowych składników determinuje umiejscowienie płatków na 3 półce. Wiele osób, które prowadzą zdrowy tryb życia patrzą też łącznie na zawartość białka, błonnika i potasu, dlatego ten model uwzględniający wszystkie te 3 zmienne wydaje się być istotny do sprawdzenia z perspektywy biznesowej i wyboru płatków przez klientów. Prawdopodobieństwo przynależności płatków do półki 3 oznaczono jako p a zmienne objaśniające (zawartość potasu, białka i błonnika odpowiednio) jako X_1, X_2, X_3.

$$\hat{L} = \ln\frac{\hat{p}}{1-\hat{p}} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -1{,}9479$	0,819	5,650 – istotny
$b_1 = 0{,}0223$	0,013	2,924 – nieistotny
$b_2 = -0{,}1265$	0,306	0,171 – nieistotny
$b_3 = 0{,}0171$	0,374	0,002 – nieistotny

Tab. 9. Tabela z parametrami dla modelu logitowego 5

$$Pseudo\ R^2 = 17{,}6\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 62{,}5\%$$
$$Precision \approx 75\%$$
$$Recall \approx 60\%$$
$$Specifity \approx 66{,}7\%$$
$$F1 \approx 66{,}7\%$$
$$AUC \approx 0{,}77$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 6	FP = 2
Przewidziana 0	FN = 4	TN = 4

Tab. 10. Macierz pomyłek modelu logitowego 5

I.5.2. Model logitowy – potas i błonnik jako zmienne objaśniające

Modelem logitowym, który zbudowano następnie jest model oparty o zmienne objaśniające potas i błonnik. Zdecydowano się na budowę tego modelu, aby sprawdzić czy tylko składniki mineralne mają wpływ na umiejscowienie płatków na półce 3. Prawdopodobieństwo przynależności płatków do półki 3 oznaczono jako p a zmienne objaśniające (zawartość potasu i błonnika odpowiednio) jako X_1, X_2.

$$\hat{L} = \ln \frac{\hat{p}}{1-\hat{p}} = b_2 X_2 + b_1 X_1 + b_0$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -2{,}1325$	0,691	9,517 – istotny
$b_1 = 0{,}0207$	0,012	2,822 – nieistotny
$b_2 = 0{,}0127$	0,371	0,001 – nieistotny

Tab. 11. Tabela z parametrami dla modelu logitowego 6

$$Pseudo\ R^2 = 17{,}4\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 62{,}5\%$$
$$Precision \approx 75\%$$
$$Recall \approx 60\%$$
$$Specifity \approx 66{,}7\%$$
$$F1 \approx 66{,}7\%$$
$$AUC \approx 0{,}78$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 6	FP = 2
Przewidziana 0	FN = 4	TN = 4

Tab. 12. Macierz pomyłek modelu logitowego 6

I.5.3. Model logitowy – potas i białko jako zmienne objaśniające

Modelem logitowym, który zbudowano następnie jest model oparty o zmienne objaśniające potas i białko. Zdecydowano się na budowę tego modelu, aby sprawdzić czy składnik budulcowy mięśni i minerał potasu wystarczą, aby płatki znalazły się na 3 półce. Prawdopodobieństwo przynależności płatków do półki 3 oznaczono jako p a zmienne objaśniające (zawartość potasu i białka odpowiednio) jako X_1, X_2.

$$\hat{L} = \ln\frac{\hat{p}}{1-\hat{p}} = b_2X_2 + b_1X_1 + b_0$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -1{,}9586$	0,785	6,220 – istotny
$b_1 = 0{,}0227$	0,008	7,690 – istotny
$b_2 = -0{,}1261$	0,306	0,170 – nieistotny

Tab. 13. Tabela z parametrami dla modelu logitowego 7

$$Pseudo\ R^2 = 17{,}6\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 62{,}5\%$$
$$Precision \approx 75\%$$
$$Recall \approx 60\%$$
$$Specifity \approx 66{,}7\%$$
$$F1 \approx 66{,}7\%$$
$$AUC \approx 0{,}75$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 6	FP = 2

Przewidziana 0	FN = 4	TN = 4

Tab. 14. Macierz pomyłek modelu logitowego 7

I.5.4. Model logitowy – białko i błonnik jako zmienne objaśniające

Modelem logitowym, który zbudowano następnie jest model oparty o zmienne objaśniające białko i błonnik. Zdecydowano się na budowę tego modelu, aby sprawdzić czy składnik budulcowy mięśni i minerał dający uczucie sytości wystarczą, aby płatki znalazły się na 3 półce. Prawdopodobieństwo przynależności płatków do półki 3 oznaczono jako p a zmienne objaśniające (zawartość białka i błonnika odpowiednio) jako X_1, X_2.

$$\hat{L} = \ln \frac{\hat{p}}{1-\hat{p}} = b_2 X_2 + b_1 X_1 + b_0$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -1{,}3914$	0,720	3,733 – nieistotny
$b_1 = 0{,}0222$	0,282	0,006 – nieistotny
$b_2 = 0{,}5595$	0,232	5,808 – istotny

Tab. 15. Tabela z parametrami dla modelu logitowego 8

$$Pseudo\ R^2 = 13{,}7\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 56{,}3\%$$
$$Precision \approx 71{,}4\%$$
$$Recall \approx 50\%$$
$$Specifity \approx 66{,}7\%$$
$$F1 \approx 58{,}8\%$$
$$AUC \approx 0{,}76$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 5	FP = 2

Przewidziana 0	FN = 5	TN = 4

Tab. 16. Macierz pomyłek modelu logitowego 8

I.5.5. Model logitowy – białko jako zmienna objaśniająca

Modelem logitowym, który zbudowano następnie jest model oparty o zmienną objaśniającą białko. Zdecydowano się na budowę tego modelu, aby sprawdzić czy wysoka zawartość składnika budulcowego mięśni wystarczy, aby płatki znalazły się na 3 półce – podyktowane jest to tym, że niektóre osoby prowadzące aktywny tryb życia, np. poprzez trenowanie siłowe, patrzą tylko na zawartość białka. Prawdopodobieństwo przynależności płatków do półki 3 oznaczono jako p a zmienną objaśniającą (zawartość białka) jako X_1.

$$\hat{L} = \ln\frac{\hat{p}}{1-\hat{p}} = b_1 X_1 + b_0$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -1{,}3279$	0,684	3,771 – nieistotny
$b_1 = 0{,}4050$	0,247	2,683 – nieistotny

Tab. 17. Tabela z parametrami dla modelu logitowego 9

$$Pseudo\ R^2 = 3{,}5\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 50\%$$
$$Precision \approx 100\%$$
$$Recall \approx 20\%$$
$$Specifity \approx 100\%$$
$$F1 \approx 33{,}3\%$$
$$AUC \approx 0{,}75$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 2	FP = 0

Przewidziana 0	FN = 8	TN = 6

Tab. 18. Macierz pomyłek modelu logitowego 9

I.5.6. Model logitowy – potas jako zmienna objaśniająca

Modelem logitowym, który zbudowano następnie jest model oparty o zmienną objaśniającą potas. Zdecydowano się na budowę tego modelu ze względu na to, że wcześniejszy model oparty o potas i białko miał najlepsze metryki w porównaniu z innymi dotychczasowymi modelami i zmienna potas jako jedyna była istotna statystycznie w tym modelu (model z błonnikiem też wykazał istotność tego składnika, ale model z potasem i białkiem miał lepsze metryki). Prawdopodobieństwo przynależności płatków do półki 3 oznaczono jako p a zmienną objaśniającą (zawartość potasu) jako X_1.

$$\hat{L} = \ln \frac{\hat{p}}{1-\hat{p}} = b_1 X_1 + b_0$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -2{,}1400$	0,656	10,634 – istotny
$b_1 = 0{,}0210$	0,007	9,211 – istotny

Tab. 19. Tabela z parametrami dla modelu logitowego 10

$$Pseudo\ R^2 = 17{,}4\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 62{,}5\%$$
$$Precision \approx 75\%$$
$$Recall \approx 60\%$$
$$Specifity \approx 66{,}7\%$$
$$F1 \approx 66{,}7\%$$
$$AUC \approx 0{,}78$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 6	FP = 2
Przewidziana 0	FN = 4	TN = 4

Tab. 20. Macierz pomyłek modelu logitowego 10

Na podstawie analizy wyników zbudowanych modeli dla półki 3, można stwierdzić, że żaden nie jest idealny. Najlepszym modelem dla półki 3 okazał się model oparty na potasie. Wykazuje on najwyższą istotność statystyczną dla zmiennej potas. Modele zawierające jednocześnie potas i błonnik charakteryzują się wysokimi wartościami p-value co może świadczyć o ich multikolinearności, co pokazuje także macierz korelacji (0,90). Model dla potasu i protein posiada wysoki AUC (~0,75, minimalnie mniejszy od modelu opartego tylko o potas) i minimalnie większą metrykę Pseudo R^2, natomiast współczynnik przy białku jest ujemny, co sugerowałoby, że im więcej białka, tym mniejsze prawdopodobieństwo wystąpienia płatków na 3 półce, co nie do końca wpasowuje się w nasze zadane pytania i macierz korelacji. Dlatego właśnie model oparty tylko o potas wydaje się być najlepszy, mimo że nie ma on idealnych metryk. Analiza wyników zbudowanych modeli pozwala stwierdzić, że na półce 3 rzeczywiście mogą znajdować się płatki zdrowsze, przede wszystkim z wysoką zawartością składnika mineralnego, jakim jest potas.

I.6. Modele dla półki 1

Macierz korelacji nie wykazała konkretnych zależności między składnikami płatków a umiejscowieniem ich na półce 1. Zdecydowano się na sprawdzenie modeli opartych o zmienne objaśniające, które z punktu widzenia klienta mogą być istotne. Dlatego zdecydowano się na zbudowanie kilku modeli dla półki 1, aby sprawdzić te zależności. Przy wyborze zmiennych objaśniających kierowano się także macierzą korelacji.

I.6.1. Model logitowy – węglowodany jako zmienna objaśniająca

Modelem logitowym, który zbudowano jako pierwszy jest model oparty o zmienną objaśniającą węglowodany. Zdecydowano się na budowę tego modelu, aby sprawdzić czy tylko większa zawartość węglowodanów decyduje o umiejscowieniu płatków na 1 półce (korelacja dodatnia). Prawdopodobieństwo przynależności płatków do półki 1 oznaczono jako p a zmienną objaśniającą (zawartość węglowodanów odpowiednio) jako X_1.

$$\hat{L} = \ln \frac{\hat{p}}{1-\hat{p}} = b_0 + b_1 X_1$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto

poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -4{,}0488$	1,391	8,474 – istotny
$b_1 = 0{,}2043$	0,085	5,774 – istotny

Tab. 21. Tabela z parametrami dla modelu logitowego 11

$$Pseudo\ R^2 = 8{,}8\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 68{,}8\%$$
$$Precision \approx 0\%$$
$$Recall \approx 0\%$$
$$Specifity \approx 78{,}6\%$$
$$F1 \approx 0\%$$
$$AUC \approx 0{,}64$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 0	FP = 2
Przewidziana 0	FN = 2	TN = 12

Tab. 22. Macierz pomyłek modelu logitowego 11

I.6.2. Model logitowy – cukier jako zmienna objaśniająca

Modelem logitowym, który zbudowano następnie jest model oparty o zmienną objaśniającą cukier. Zdecydowano się na budowę tego modelu, aby sprawdzić czy tylko mniejsza zawartość cukru (korelacja ujemna) ma wpływ na umiejscowienie płatków na półce 1. Prawdopodobieństwo przynależności płatków do półki 3 oznaczono jako p a zmienną objaśniającą (zawartość cukru odpowiednio) jako X_1.

$$\hat{L} = \ln \frac{\hat{p}}{1-\hat{p}} = b_1 X_1 + b_0$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = 0{,}1652$	0,502	0,108 – nieistotny
$b_1 = -0{,}1724$	0,075	5,235 – istotny

Tab. 23. Tabela z parametrami dla modelu logitowego 12

$$Pseudo\ R^2 = 8{,}2\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 75\%$$
$$Precision \approx 0\%$$
$$Recall \approx 0\%$$
$$Specifity \approx 85{,}7\%$$
$$F1 \approx 0\%$$
$$AUC \approx 0{,}61$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 0	FP = 2
Przewidziana 0	FN = 2	TN = 12

Tab. 24. Macierz pomyłek modelu logitowego 12

I.6.3. Model logitowy – węglowodany i cukier jako zmienne objaśniające

Modelem logitowym, który zbudowano następnie jest model oparty o zmienne objaśniające węglowodany i cukier. Zdecydowano się na budowę tego modelu, aby sprawdzić czy większa zawartość węglowodanów i mniejsza zawartość cukru wpływają na umiejscowienie danych płatków na półce 1. Prawdopodobieństwo przynależności płatków do półki 1 oznaczono jako p a zmienne objaśniające (zawartość węglowodanów i cukru odpowiednio) jako X_1, X_2.

$$\hat{L} = \ln\frac{\hat{p}}{1-\hat{p}} = b_2X_2 + b_1X_1 + b_0$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$

$b_0 = -2{,}2835$	1,874	1,484 – nieistotny
$b_1 = 0{,}1333$	0,097	1,874 – nieistotny
$b_2 = -0{,}1098$	0,088	1,545 – nieistotny

Tab. 25. Tabela z parametrami dla modelu logitowego 13

$$Pseudo\ R^2 = 11\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 75\%$$
$$Precision \approx 0\%$$
$$Recall \approx 0\%$$
$$Specifity \approx 85{,}7\%$$
$$F1 \approx 0\%$$
$$AUC \approx 0{,}71$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 0	FP = 2
Przewidziana 0	FN = 2	TN = 12

Tab. 26. Macierz pomyłek modelu logitowego 13

I.6.4. Model logitowy – wszystkie składniki płatków jako zmienne objaśniające

Modelem logitowym, który zbudowano następnie jest model oparty o zmienne objaśniające będące wszystkimi składnikami płątków. Zdecydowano się na budowę tego modelu, aby sprawdzić czy może cały skład ma wpływ na umiejscowienie płatków na 1 półce. Prawdopodobieństwo przynależności płatków do półki 1 oznaczono jako p a zmienne objaśniające (zawartość kalorii, cukrów, węglowodanów, białka, tłuszczu, sodu, błonnika i potasu odpowiednio) jako X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8.

$$\hat{L} = \ln\frac{\hat{p}}{1-\hat{p}} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -2{,}0621$	3,248	0,403 – nieistotny
$b_1 = -0{,}0552$	0,063	0,762 – nieistotny
$b_2 = 0{,}1592$	0,291	0,299 – nieistotny
$b_3 = 0{,}3080$	0,300	1,055 – nieistotny
$b_4 = 0{,}6705$	0,451	2,211 – nieistotny
$b_5 = -0{,}2312$	0,637	0,132 – nieistotny
$b_6 = 0{,}0020$	0,004	0,270 – nieistotny
$b_7 = 0{,}0745$	0,406	0,033 – nieistotny
$b_8 = -0{,}0095$	0,015	0,392 – nieistotny

Tab. 27. Tabela z parametrami dla modelu logitowego 14

$$Pseudo\ R^2 = 17\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 75\%$$
$$Precision \approx 0\%$$
$$Recall \approx 0\%$$
$$Specifity \approx 85{,}7\%$$
$$F1 \approx 0\%$$
$$AUC \approx 0{,}79$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 0	FP = 2
Przewidziana 0	FN = 2	TN = 12

Tab. 28. Macierz pomyłek modelu logitowego 14

Na podstawie przeprowadzonych obliczeń oraz wcześniejszej analizy macierzy korelacji, należy stwierdzić, że nie udało się zbudować modelu o wysokiej zdolności klasyfikacyjnej dla półki 1. Bardzo niskie (równe 0) wartości metryk precyzji, czyłości i F1 oraz nieistotność statystyczna większości zmiennych objaśniających (poza węglowodanami i cukrem i modelach, które zawierają tylko je) w modelach

wielu zmiennych dowodzą, że w badanym zbiorze danych nie istnieją silne, powtarzalne zależności między składem odżywczym płatków a ich umiejscowieniem na dolnej półce. Potwierdza to wstępną analizę macierzy korelacji, która nie wykazała żadnego składnika wyraźnie powiązanego z tą półką. Słabe wyniki modeli sugerują, że o lokowaniu płatków na samym dole regału mogą decydować czynniki inne niż wartości odżywcze. Najprawdopodobniej są to kryteria logistyczne, np. gabaryty opakowań lub ekonomiczne, takie jak niska cena.

I.7. Wyłączenie wybranych 2 producentów płatków dla każdej z półek 2-3 w celu weryfikacji poprawności klasyfikacji

Zrezygnowano z uwzględnienia półki 1 w poniższych rozważaniach ze względu na brak wykrytych jednoznacznych zależności między składem a umiejscowieniem na półce 1. Zdecydowano się na wyłączenie z danych producenta z największą liczbą płatków w zbiorze danych (Producent K – 23 płatki), aby mieć na czym weryfikować działanie modeli.

I.7.1. Najlepszy model dla półki 3 – oparty o potas

Model logitowy był już zbudowany wyżej - I.3 dla potasu. Model wytrenowano na danych nie uwzględniających producenta „K".

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -1{,}9984$	0,680	8,638 – istotny
$b_1 = 0{,}0192$	0,007	8,111 – istotny

Tab. 29. Tabela z parametrami dla modelu logitowego półki 3 po wyłączeniu producenta

$$Pseudo\ R^2 = 16{,}1\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 74\%$$
$$Precision \approx 87{,}5\%$$
$$Recall \approx 58{,}3\%$$
$$Specifity \approx 90{,}9\%$$
$$F1 \approx 70\%$$
$$AUC \approx 0{,}83$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 7	FP = 1
Przewidziana 0	FN = 5	TN = 10

Tab. 30. Macierz pomyłek modelu logitowego półki 3 po wyłączeniu producenta

I.7.2. Najlepszy model dla półki 3 – oparty o potas po optymalizacji progu klasyfikacji

Wykorzystano ten sam model, ale z optymalnym progiem klasyfikacji.

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$. Parametry, ich odchylenia i statystyki Walda są identyczne jak dla poprzedniego modelu.

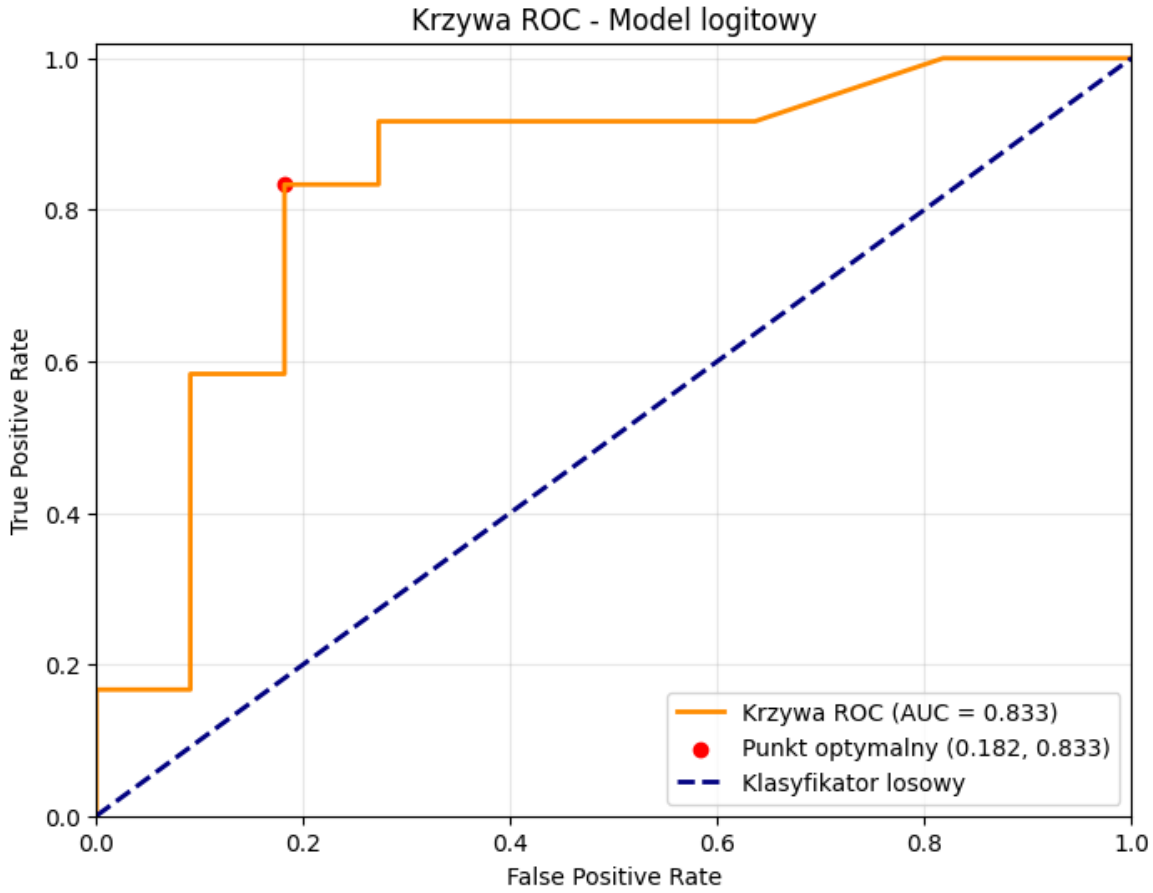
$$Pseudo\ R^2 = 16{,}1\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 82{,}6\%$$
$$Precision \approx 83{,}3\%$$
$$Recall \approx 83{,}3\%$$
$$Specifity \approx 81{,}8\%$$
$$F1 \approx 83{,}3\%$$
$$AUC \approx 0{,}83$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 10	FP = 2
Przewidziana 0	FN = 2	TN = 9

Tab. 31. Macierz pomyłek modelu logitowego półki 3 po wyłączeniu producenta i z optymalnym progiem klasyfikacji



Rys. 2. Krzywa ROC dla modelu logitowego półki 3

Po wyłączeniu z danych treningowych płatków producenta „K", który miał najwięcej płatków w naszym zbiorze początkowych danych, zbudowane modele całkiem dobrze sobie radzą z klasyfikacja płatków tego producenta na półce 3. Zmienne w obu modelach są istotne statystycznie, zgodnie z tym co opisano we wstępie teoretycznym - I.1.2. Dla pierwszego modelu, zdolność do klasyfikacji płatków na półce 3 charakteryzuje się wyższą zdolnością do klasyfikacji wykluczającej należenie do półki (specyficznością) jednakże wartość F1 jest na poziomie ok. 70%, co jest całkiem dobrym wynikiem. Należy jednak wziąć pod uwagę to, że w kontekście biznesowym, zależy nam na jak największej jakości klasyfikacji płatków jako należących do półki 3, dlatego ten model nie jest najlepszy jeżeli weźmiemy pod uwagę kontekst biznesowy.

W modelach logitowych wartości $Pseudo\ R^2$ rzędu 0,2-0,4 są uznawane za bardzo dobre dopasowanie. Wynik w okolicach 0,16, przy tak małej próbie i tylko jednej zmiennej objaśniającej należy zatem uznać za satysfakcjonujący i potwierdzający istotność potasu jako czynnika klasyfikującego płatki na półkę 3.

Dzięki zastosowaniu optymalnego progu klasyfikacji wyznaczonego metodą krzywej ROC, udało się polepszyć wyniki metryk oceny jakości klasyfikacji modelu, m.in. zwiększyć czułość modelu o ok. 25 punktów procentowych. Oznacza to, że model znacznie lepiej radzi sobie z wykrywaniem produktów, które faktycznie powinny znaleźć się na półce 3 (a na tym, jak już wspomniano, nam zależy), kosztem jedynie niewielkiego spadku specyficzności. Należy jednak zauważyć, że specyficzność i czułość są zbliżone do siebie a wartość F1 jest dość wysoka, porównując ją chociażby z poprzednim modelem, co również świadczy o poprawności skuteczności modelu.

Podsumowując, modele wykazują wysoką zdolność do generalizacji po wyłączeniu kluczowego producenta ze zbioru treningowego, w szczególności jeśli mówimy o modelu z optymalnym progiem klasyfikacji. Model wciąż poprawnie klasyfikuje większość danych dla tego producenta. Można również przypuszczać, że zastosowanie krzywej ROC w celu wyboru optymalnego progu klasyfikacji, poprawiłoby także jakość poprzednich modeli, trenowanych na całym zbiorze danych.

I.7.3. Najlepszy model dla półki 2 – oparty o cukier

Model logitowy był już zbudowany wyżej - I.1.1 dla cukru. Model wytrenowano na danych nie uwzględniających producenta „K".

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$.

Parametr b_i	Odchylenie S(b_i)	$W_i > \chi^2_{kryt}$
$b_0 = -2{,}0909$	0,692	9,139 – istotny
$b_1 = 0{,}1426$	0,077	3,389 – nieistotny

Tab. 32. Tabela z parametrami dla modelu logitowego półki 2 po wyłączeniu producenta

$$Pseudo\ R^2 = 6\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 73{,}9\%$$
$$Precision \approx 100\%$$
$$Recall \approx 14{,}3\%$$

$$Specifity \approx 100\%$$
$$F1 \approx 25\%$$
$$AUC \approx 0{,}89$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 1	FP = 0
Przewidziana 0	FN = 6	TN = 16

Tab. 33. Macierz pomyłek modelu logitowego półki 2 po wyłączeniu producenta

I.7.4. Najlepszy model dla półki 2 – oparty o cukier po optymalizacji progu klasyfikacji

Wykorzystano ten sam model, ale z optymalnym progiem klasyfikacji.

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie α = 0,05. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3{,}841$. Parametry, ich odchylenia i statystyki Walda są identyczne jak dla poprzedniego modelu.

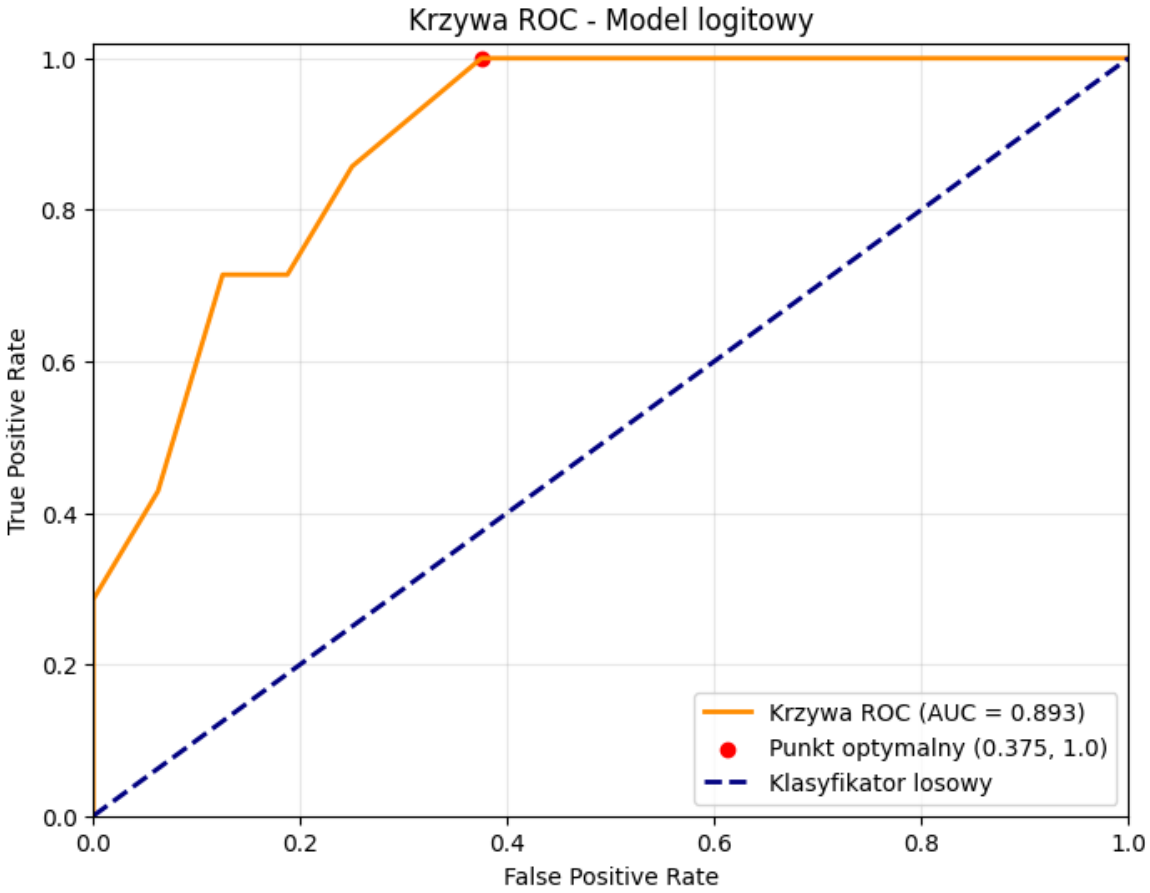
$$Pseudo\ R^2 = 6\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 73{,}9\%$$
$$Precision \approx 53{,}8\%$$
$$Recall \approx 100\%$$
$$Specifity \approx 62{,}5\%$$
$$F1 \approx 70\%$$
$$AUC \approx 0{,}89$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 7	FP = 6
Przewidziana 0	FN = 0	TN = 10

Tab. 34. Macierz pomyłek modelu logitowego półki 2 po wyłączeniu producenta i z optymalnym progiem klasyfikacji



Rys. 3. Krzywa ROC dla modelu logitowego półki 2

Model logitowy dla półki 2 oparte na zmiennej objaśniającej cukier przed zmianą progu klasyfikacji, wykazuje znacznie słabsze właściwości klasyfikacyjne niż analogiczne modele dla półki 3. Niska wartość $Pseudo\ R^2$ (poniżej 0,1) oraz niska czułość sugerują, że model wytrenowany na danych z wyłączeniem jednego producenta nie dopasowuje się najlepiej do danych oraz nie ma wysokiej jakości klasyfikacji płatków jako należących do półki 2 – nie doszacowuje prawdopodobieństwa przynależności płatków tego producenta do tej półki . Model ma tendencję do zaniżania prawdopodobieństwa przynależności płatków do półki 2. Dowodzi to też temu, że model de facto wyuczony na konkurencji nie potrafi w pełni zrekonstruować zależności które odkryto przy pełnych danych.

Po zmianie progu klasyfikacji dla półki 2 zauważono znaczącą poprawę jakości modelu. Niskie wartości czułości czy F1 sugerowały, że model logitowy poprawnie wyklucza płatki według zawartości cukru, ale standardowy próg 0,5 nie był optymalny.

Po zmianie progu na 0,25 model usiągnął pełną czułość (100%) przy wysokim wskaźniku F1 (70%). Podczas gdy model wyuczony na innych płatkach innych

producentów oczekuje wysokich wartości cukru dla 2. Półki, producent K umieszcza tam płatki już przy niższych zawartościach cukru. Optymalizacja progu pozwoliła na poprawę jakości klasyfikacji modelu jednakże nie jest to też najlepszy model, m.in. poprzez małą licznbość zbioru danych, co po wyłączeniu producenta sprawia, że model jest gorszy od poprzednich i zmienna cukier jest nieistotna statystycznie czy chociażby sama 100% wartość czułości przy jednoczesnym precyzji może sugerować, że ten konkretny model klasyfikuje na półkę 2 niemal każdy produkt o nawet umiarkowanej zawartości cukru.

Wnioski

W wyniku analizy zbudowanych modeli stwierdzono, że skład odżywczy płatków śniadaniowych ma wpływ na ich umiejscowienie na danej półce, ale w sposób niejednoznaczny. Najsilniejsze zależności zaobserwowano dla półek 3 i 2, podczas gdy dla półki 1 nie stwierdzono jednoznacznych zależności determinujących umiejscowienie płatków na tej półce. Analiza potwierdziła, że na półce 2, znajdującej się na wysokości wzroku, ulokowane są płatki o gorszej jakości, co wykazał model oparty o zmienną objaśniającą – cukier. Istotność statystyczna tej zmiennej oraz jej dodatnia korelacja wskazują na możliwą celową strategię sklepu polegającą na eksponowaniu na tej półce płatków o wysokiej zawartości cukry. Z kolei półka 3 zidentyfikowana została jako zawierająca płatki o zdrowszym składzie, gdzie najlepszym czynnikiem decydującym o umiejscowieniu płatków na tej półce okazała się zawartość potasu, która jest silnie skorelowana z błonnikiem. W przypadku modeli wielu zmiennych dla tej półki konieczna była redukcja zmiennych ze względu na możliwą multikolinearność między składnikami potasu i błonnika. Odmienne rezultaty uzyskano dla półki 1, dla której nie udało się zbudować modelu o zadowalającej jakości klasyfikacyjnej. Modele logitowe dla tej półki uzyskały zerowe wartości precyzji i czułości (oraz F1) w ujęciu modeli wielu zmiennych, dodatkowo wykazano nieistotność większości zmiennych objaśniających w zbudowanych modelach, co świadczy brak jednoznacznego związku między składem odżywczym płatków a umiejscowieniem ich na tej półce. Oznacza to, że o umieszczeniu płatków na najniższej półce decydują czynniki inne niż te żywieniowe. Podsumowując, modele logitowe skutecznie wykryły podział płatków na swego rodzaju strefy „impulsywną", w której są płatki wysokosłodzone, których sklep szybko chce się pozbyć, oraz „zdrową", gdzie są produkty o zwiększonej zawartości zdrowych składników. Modele okazały się jednak niewystarczające, aby wyjaśnić strukturę półki 1, która nie zależy od składników odżywczych.

Przeprowadzone eksperymenty polegające na wyłączeniu z danych treningowych danych największego producenta – „K", pozwoliły na ocenę zdolności klasyfikacyjnych modeli wobec nowych, nieznanych wcześniej danych. Dla póki 3, modele wykazują wysoką zdolność do generalizacji zasad dot. umiejscowienia płatków na półce 3 według składu. Nawet po wyłączeniu producenta K, model oparty na potasie poprawnie klasyfikuje jego produkty na najwyższej półce (osiągając F1 na poziomie 70% czy 83,3% po optymalizacji progu klasyfikacji). Dowodzi to temu, że na półce 3 rzeczywiście są umiejscawiane płatki z największą zawartością potasu. Dla półki 2, eksperyment wykazał, że można się spodziewać, że zasady umiejscawiania płatków na tej półce są bardziej zróżnicowane i zależne od konkretnego producenta. Modele wytrenowane bez uwzględnienia danych największego producenta okazały się nie doszacowywać prawdopodobieństwa. Dopiero

optymalizacja progu klasyfikacji do poziomu 0,25 pozwoliła na uzyskanie zadowalającego wyniku F1 oraz bardzo wysokiej czułości. Warto jednak zauważyć, że dla półki 2., ze względu na małą liczność danych, zmienna objaśniająca cukier stała się nieistotna statystycznie. Wykazano, że domyślny próg klasyfikacji na poziomie 0,5 jest często nieefektywny dla tego problemu biznesowego. Zastosowanie krzywej ROC do wyznaczenia optymalnego progu klasyfikacji pozwoliło na znaczącą poprawę jakości klasyfikacji, szczególnie w przypadku półki 2.

Bibliografia

[1] https://en.wikipedia.org/wiki/Pseudo-R-squared
[2] https://miroslawmamczur.pl/jak-dziala-regresja-logistyczna/
[3] https://www.youtube.com/watch?v=idGS1gLIXTA
[4] https://www.youtube.com/watch?v=2HpcV1sq0fE
[5] https://en.wikipedia.org/wiki/Pseudo-R-squared
[6] https://pl.wikipedia.org/wiki/Funkcja_wiarygodno%C5%9Bci
[7] https://www.naukowiec.org/wzory/statystyka/statystyka-walda--test_455.html
[8] https://stats.stackexchange.com/questions/82105/mcfaddens-pseudo-r2-interpretation

Spis rysunków

Spis tabel

Załączniki

Lab-2-Zad-3-Michał-Ślęzak-Szymon-Oleśkiewicz.ipynb

Lab-2-Zad-3-Michał-Ślęzak-Szymon-Oleśkiewicz.py

https://colab.research.google.com/drive/1aUw_1s-62GefoAwS-hrYgUTSez5aw3GB?usp=sharing