ZESZYT NAUKOWY
Wyższa Szkoła Zarządzania i Bankowości w Krakowie

dr Anna Kijanka
Wyższa Szkoła Zarządzania i Bankowości w Krakowie
a.kijanka@wszib.edu.pl
mgr Tomasz Molęcki
Wyższa Szkoła Zarządzania i Bankowości w Krakowie
t.molecki@wszib.edu.pl

ZASTOSOWANIE MODELU LOGITOWEGO DO PROGNOZOWANIA LICZBY STUDENTÓW KONTYNUUJĄCYCH KSZTAŁCENIE NA STUDIACH II STOPNIA W MACIERZYSTEJ UCZELNI NIEPUBLICZNEJ

Wprowadzenie

Podstawowym źródłem przychodów niepublicznych uczelni wyższych są wnoszone przez studentów opłaty, głównie w postaci czesnego. Zarządzający uczelniami stają przed istotnym wyzwaniem, polegającym na zapewnieniu przychodów z opłat studentów na takim poziomie, który pozwoli na sfinansowanie kosztów operacyjnych oraz wygenerowanie dodatniego wyniku finansowego dla instytucji. Wysoce pożądane jest więc, aby na studia zgłosiła się odpowiednia liczba kandydatów, dzięki której możliwe będzie zagwarantowanie właściwego poziomu finansów uczelni. W szczególności ważna wydaje się być dla zarządzających uczelnią wiedza dotycząca wyboru dalszej ścieżki kariery przez absolwentów studiów I stopnia, pozwalająca na podjęcie takich działań, by jak największa ich liczba zdecydowała się na kontynuację nauki na studiach uzupełniających magisterskich (SUM) w macierzystej uczelni.

Kwerenda literatury przedmiotu wskazuje na pewne zainteresowanie przedstawionym problemem, jednak liczba dostępnych publikacji jest niewielka[1]. Autorzy istniejących opracowań poświęcają głównie uwagę kwestiom determinant jakości[2] oraz kosztów kształcenia[3].

Celem niniejszego opracowania jest oszacowanie prawdopodobieństwa kontynuacji kształcenia na uzupełniających studiach magisterskich w macierzystej uczelni, a tym samym przybliżenie liczby kandydatów na studia II stopnia, rekrutujących się z grona dotychczasowych absolwentów studiów licencjackich.

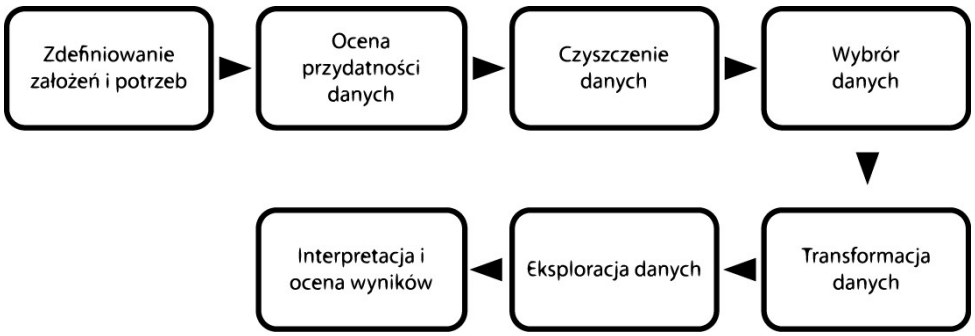
[1] A. Nandeshwar S. Chaudhari *Enrollment Prediction Models Using Data Mining,* http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf (data odczytu 22.05.2022).
[2] G. Zieliński, K. Lewandowski, *Determinanty percepcji jakości usług edukacyjnych w perspektywie grup interesariuszy*, http://zif.wzr.pl/pim/2012_3_3_4.pdf (data odczytu 25.05.2022).
[3] I. Gawryś, P. Trippner, *Analiza poziomu rentowności przedsiębiorstwa na przykładzie niepublicznej uczelni wyższej w roku akademickim 2015/2016*, „Annales Universitatis Mariae Curie-Skłodowska", 2017, nr LI, 5.

1. Odkrywanie wiedzy

Dane stanowiące materiał badawczy zostały pobrane z systemu zarządzania uczelnią w jednej z krakowskich niepublicznych szkół wyższych. Wybór danych podyktowany został przede wszystkim ich dostępnością. Wynikał również z subiektywnej oceny autorów opracowania w kwestii ich przydatności w realizacji celu pracy.

Określenie informacji (zmiennych), które zostały wykorzystane do analiz wymagało podjęcia wielu czynności, w szczególności: czyszczenia danych (poprawy błędów, uzupełnienia brakujących danych), transformacji danych (zakodowania lub usunięcia danych osobowych studentów, wzbogacenia danych poprzez obliczanie nowych atrybutów, agregacji danych). Wymienione działania stanowiły wstęp do procesu *eksploracji danych* (rysunek 1), który pozwala na odkrywanie w danych asocjacji, ukrytych wzorców, reguł oraz trendów i w efekcie pozwala na odkrywanie wiedzy[4].

Rysunek 1. Etapy procesu odkrywania wiedzy



Źródło: opracowanie własne na podstawie: M. Szeliga, *Data Science i uczenie maszynowe*, Wydawnictwo Naukowe PWN, Warszawa 2017, s. 2-3.

Szczegółowy opis zmiennych poddanych analizie zestawiono w tabeli 1.

Zmienna zależna „SUM" przedstawia zerojedynkową, zakodowaną informację, określającą, czy absolwent studiów I stopnia podjął naukę na studiach uzupełniających magisterskich w macierzystej uczelni. Wyodrębniono następujące zmienne objaśniające: „Płeć", „Obrona rok", „Forma studiów", „Średnia ocen", „Czy otrzymał stypendium", „Wiek obrona", „Odległość", których nazwy ze względu na wymogi oprogramowania wykorzystywanego w procesie modelowania zostały uproszczone odpowiednio do: *plec*, *obrona_rok*, *tryb*, *srednia*, *stypendium*, *wiek, km*.

[4] T. Morzy, *Eksploracja danych*, Wydawnictwo Naukowe PWN, Warszawa 2013, s. 3-9.

Tabela 1. Zmienne poddane procesowi eksploracji

Nazwa zmiennej	Opis
SUM (*sum*)	Zmienna jakościowa określająca, czy student kontynuował naukę na studiach uzupełniających magisterskich. Informacja zakodowana w postaci 0 (brak kontynuacji) oraz 1 (kontynuacja).
Płeć (*plec*)	Zmienna jakościowa opisująca płeć, zakodowana w postaci: 0 (mężczyzna), 1 (kobieta).
Obrona rok (*obrona_rok*)	Rok obrony pracy licencjackiej (zmienna ilościowa).
Forma studiów (*tryb*)	Zmienna jakościowa opisująca tryb studiów. Zmienna została zakodowana w formie 0 (studia stacjonarne) oraz 1 (studia niestacjonarne).
Średnia ocen (*srednia*)	Średnia ocen ze studiów, przedstawiona w skali od 2 do 5. Ocena wyliczona na podstawie ocen semestralnych.
Stypendium (*stypendium*)	Informacja o otrzymaniu przez studenta stypendium. Zmienna jakościowa zakodowana w postaci 0 (brak stypendium) oraz 1 (stypendium).
Wiek obrona (*wiek*)	Zmienna wyliczeniowa reprezentująca wiek studenta w roku obrony pracy licencjackiej. Podstawą do obliczenia zmiennej jest rok obrony pracy oraz data urodzenia studenta.
Odległość (*km*)	Wyliczeniowa zmienna ilościowa zawierająca odległość miejsca zamieszkania studenta do Krakowa, podana w kilometrach.

Źródło: opracowanie własne.

Do realizacji celu badania zastosowano metodę eksploracji danych z wykorzystaniem modelu regresji logistycznej. Wybór metody badawczej podyktowany został jakościowym charakterem zmiennej objaśnianej. Oczekiwano, że zbudowany model pozwoli na wskazanie czynników determinujących decyzje absolwentów studiów I stopnia o dalszym kształceniu na uzupełniających studiach magisterskich w macierzystej uczelni.

Stworzono 4 odrębne modele ekonometryczne dla poszczególnych kierunków, na których uczelnia prowadzi kształcenie. W niniejszym artykule zaprezentowano model zbudowany na podstawie danych dotyczących studentów i absolwentów kierunku *Zarządzanie*. Wykorzystano informacje o absolwentach, którzy uzyskali dyplom w latach 2013-2018.

2. Istota modelu regresji logistycznej

Model regresji logistycznej, określany też jako model logitowy lub model prawdopodobieństwa, pozwala na określenie prawdopodobieństwa przynależności obiektu do jednej z dwóch klas, w zależności od charakteryzującego go wektora zmiennych niezależnych[5]. Charakterystyczna jest dla regresji logistycznej właśnie binarna, jakościowa zmienna objaśniana.

[5] T. Kufel, *Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL,* Wydawnictwo Naukowe PWN, Warszawa 2011, s. 142.

Zmienna zależna reprezentuje najczęściej wystąpienie lub brak wystąpienia zdarzenia, które chce się prognozować, np.: stan upadłości firmy (1 – tak, 0 – nie), podjęcie kształcenia na danym poziomie (1 – tak, 0 – nie). Ponieważ model regresji liniowej, bez narzucania na niego dodatkowych warunków, nie jest w stanie zapewnić, aby wartość zmiennej zależnej należała zawsze do przedziału odpowiadającego możliwym prawdopodobieństwom, do modelowania prawdopodobieństwa wykorzystuje się model postaci:

$$P = P(Y = 1/X_1, X_2, \ldots, X_k) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k)}} \qquad (1)$$

gdzie:

- Y jest binarną zmienną zależną (objaśnianą),
- $X_1, X_2, \ldots, X_k$ to k zmiennych niezależnych (objaśniających), które mogą być mierzalne lub jakościowe,
- P oznacza prawdopodobieństwo warunkowe sukcesu (przynależności obiektu do kategorii kodowanej jako 1),
- α_0, α_1, …, α_k to parametry (współczynniki) strukturalne modelu[6].

W wyniku prostego przekształcenia zależności (1) otrzymuje się równoważną postać modelu regresji logistycznej:

$$\frac{P}{1 - P} = e^{\alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k} \qquad (2)$$

gdzie lewa strona równania oznacza *szansę* zajścia sukcesu, czyli stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki.

Po zlogarytmowaniu obu stron równania (2) uzyskuje się trzecią, równoważną, *logitową* postać modelu:

$$logit\ P = \ln \frac{P}{1 - P} = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k \qquad (3)$$

Najpowszechniej stosowaną metodą szacowania parametrów modelu regresji logistycznej jest metoda największej wiarygodności, ponieważ estymatory uzyskane za jej pomocą są zgodne, asymptotycznie nieobciążone, asymptotycznie efektywne, mają asymptotyczny rozkład normalny[7].

Uzyskane oceny parametrów modelu regresji logistycznej mają następującą interpretację:

[6] A. Stanisz, *Modele regresji logistycznej,* wyd. StatSoft Polska, Kraków 2016, s. 166.
[7] Ibidem, s. 63.

- jeżeli $\alpha_i > 0$, to uznać należy, że wzrost wartości cechy X_i, przy niezmienności pozostałych zmiennych, prowadzi do wzrostu prawdopodobieństwa sukcesu tzn. przynależności obiektu do klasy kodowanej jako 1,
- $\alpha_i < 0$ oznacza, że wzrost wartości zmiennej X_i, *ceteris paribus*, prowadzi do spadku prawdopodobieństwa sukcesu.

Wnioski odnośnie modelowanego zjawiska można wyrazić również w terminach *szans.* Relatywną zmianę możliwości wystąpienia zdarzenia pod wpływem czynnika opisanego przez zmienną X_i określa e^{α_i}:

- jeżeli $e^{\alpha_i} > 1$, to czynnik opisywany przez zmienną X_i ma stymulujący wpływ na wystąpienia badanego zjawiska,
- jeżeli $e^{\alpha_i} < 1$, to czynnik X_i działa hamująco,

przy czym wzrost zmiennej X_i o jednostkę skutkuje zmianą o $(e^{\alpha_i} - 1) \cdot 100\%$ szansy przynależności obiektu do klasy kodowanej jako 1 przy niezmiennym wpływie pozostałych zmiennych niezależnych.

Oszacowane współczynniki modelu α_i interpretowane w terminach *logitu* informują o zmianie wartości logarytmu szansy związanej ze zmianą o jednostkę czynnika opisanego przez X_i. Kiedy prawdopodobieństwa są większe (mniejsze) od 0,5 to wartości logit są większe (mniejsze) od 0.

Zaletą regresji logistycznej jest to, że nie wymaga niektórych założeń koniecznych dla regresji liniowej. Wektor zmiennych objaśniających i reszty nie muszą mieć rozkładu normalnego, dopuszczalna jest heteroskedastyczność. Konieczne jest jednak spełnienie kilku innych warunków:

- Zależność między logarytmem szans a wektorem zmiennych objaśniających musi być liniowa, zgodnie z równaniem (3).
- Obserwacje muszą być niezależne.
- Model powinien uwzględniać wszystkie istotne zmienne.
- Zmienne niezależne nie mogą być współliniowe.
- Regresja logistyczna jest wrażliwa na występowanie punktów odstających. Przed rozpoczęciem analizy należy je usunąć.

Podsumowując charakterystykę modelu regresji logistycznej należy podkreślić jego zasadniczą cechę, związaną z możliwością kwantyfikacji prawdopodobieństwa przynależności określonej obserwacji do klasy kodowanej jako 1. Dzięki temu możliwe jest także zaszeregowanie

(klasyfikacja) tego obiektu do jednej z dwóch klas. Reguła klasyfikacyjna[8] zwykle przyjmuje wartość „odcinającą" prawdopodobieństwa $p = 0,5$.

3. Model prawdopodobieństwa kontynuacji kształcenia na poziomie SUM

Model logitowy do oszacowania prawdopodobieństwa podjęcia nauki na SUM w macierzystej uczelni przez absolwenta kierunku *Zarządzanie* powstał w wyniku szeregu sformalizowanych czynności[9]. Wartości parametrów regresji zostały wyznaczone na podstawie danych dotyczących 1 205 absolwentów (próby uczącej) z lat 2013-2017. Dla roku 2018 (próba weryfikująca – 144 absolwentów) przeprowadzono prognozę *ex post*, aby porównać jej wyniki z danymi empirycznymi i sprawdzić w ten sposób możliwości prognostyczne modelu.

Etap estymacji parametrów modelu poprzedzono podstawową analizą struktury danych.

Podstawowe miary opisu statystycznego zmiennych

Niespełna 38% absolwentów studiów I stopnia na kierunku *Zarządzanie*, którzy ukończyli studia w latach 2013-2017, podjęło decyzję o kontynuacji studiów na SUM.

W tabeli 2 zestawiono podstawowe miary opisu statystycznego objaśniających zmiennych ilościowych: *srednia*, *wiek*, *km.*

Tabela 2. Miary położenia i zróżnicowania zmiennych (próba ucząca)

Miary	Zmienne		
	srednia	*wiek*	*km*
Średnia	4,07	28,12	44,84
Mediana	4,04	25	26,00
Odchylenie standardowe	0,373	6,883	88,265
Minimum	3,16	21	0
Maksimum	4,98	62	1000
Liczność	1205	1205	1205
Współczynnik zmienności	9%	24%	197%

Źródło: opracowanie i obliczenia własne.

Połowa absolwentów badanego kierunku, którzy otrzymali dyplom w latach 2013-2017, uzyskała średnią ocen nie mniejszą niż 4,04. Zwraca uwagę niewielkie zróżnicowanie średnich

[8] G. G. Judge, C. Hill, W. E. Griffiths, H.Lütkepohl, T. Lee, *The Theory and Practice of Econometrics*, John Wiley&Sons, New York 1985.
[9] G. James, D. Witten, T. Hastie, R. Tibshira, *An Introduction to Statistical Learning with Applications in R*, Springer, New York 2013, s. 131-135.

ocen (współczynnik zmienności na poziomie 9%). Średnia wieku absolwentów kierunku *Zarządzanie* to niewiele ponad 28 lat, przy czym najstarszy absolwent uzyskał tytuł licencjata w wieku 62 lat. Miejsce zamieszkania połowy badanych absolwentów znajdowało się nie dalej niż 26 km od Krakowa, przy średniej odległości zamieszkania od uczelni na poziomie 44 km. Zmienną *km* charakteryzuje wysoki współczynnik zmienności (197%)[10].

Wyznaczono udziały procentowe poszczególnych kategorii w zmiennych jakościowych: *plec*, *obrona_rok*, *tryb*, *stypendium* oraz *sum*. Proporcje płci w analizowanej próbie są w miarę wyrównane, kobiety stanowią 55,8% próby. Ponad 83% absolwentów studiowało w trybie niestacjonarnym. Stypendium otrzymywało niewiele ponad 39% studentów.

Wyznaczone współczynniki korelacji liniowej Pearsona (tabela 3) wskazują na przeciętną korelację pomiędzy zmiennymi *stypendium* oraz *srednia* (R=0,51) *srednia* oraz *plec* (R=0,34), a także *wiek* oraz *srednia* (R=0,36). Siła pozostałych związków liniowych jest nikła ($|R|$<0,3).

Tabela 3. Tabela korelacji pomiędzy zmiennymi objaśniającymi

Zmienna	*plec*	*obrona_rok*	*tryb*	*srednia*	*stypendium*	*wiek*	*km*
plec	1	0,00	0,05	0,34	0,25	0,05	0,01
obrona_rok	0,00	1	0,01	0,07	-0,02	0,04	0,10
tryb	0,05	0,01	1	0,12	0,05	0,29	-0,22
srednia	0,34	0,07	0,12	1	0,51	0,36	-0,03
stypendium	0,25	-0,02	0,05	0,51	1	0,14	-0,04
wiek	0,05	0,04	0,29	0,36	0,14	1	-0,16
km	0,01	0,10	-0,22	-0,03	-0,04	-0,16	1

Źródło: opracowanie i obliczenia własne.

Estymacja parametrów modelu

Parametry modelu regresji logistycznej oszacowano za pomocą programu *Gretl* (tabela 4).

[10] Wartości odstające nie mają istotnego wpływu na zdolności predykcyjne modelu, co zostało zweryfikowane poprzez eliminację wartości odstających i porównanie wyników predykcji modelu.

Tabela 4. Estymacja Logit, wykorzystane obserwacje 1-1205, Zmienna zależna (Y): sum

	Współczynnik	*Błąd stand.*	*z*	*wartość p*	
const	117,906	90,9627	1,296	0,1949	
plec	−0,0937891	0,131848	−0,7113	0,4769	
obrona_rok	−0,0600870	0,0451917	−1,330	0,1836	
tryb	−0,287385	0,173286	−1,658	0,0972	*
srednia	0,550652	0,209129	2,633	0,0085	***
stypendium	0,503834	0,142989	3,524	0,0004	***
wiek	0,0161146	0,00970125	1,661	0,0967	*
km	0,000299949	0,000701698	0,4275	0,6690	

Średn.aryt.zm.zależnej	0,375104	Odch.stand.zm.zależnej	0,484351
McFadden R-kwadrat	0,331648	Skorygowany R-kwadrat	0,221613
Logarytm wiarygodności	−772,0163	Kryt. inform. Akaike'a	1560,033
Kryt. bayes. Schwarza	1600,786	Kryt. Hannana-Quinna	1575,381

Źródło: opracowanie i obliczenia własne przy wykorzystaniu aplikacji *Gretl*.

Przyjęto poziom istotności $\alpha = 10\%$. Eliminując sekwencyjnie nieistotne statystycznie zmienne (*wartość p* > 0,1) otrzymano model o parametrach zaprezentowanych w tabeli 5.

Tabela 5. Model 2: Estymacja Logit, wykorzystane obserwacje 1-1205, Zmienna zależna (Y): sum

	Współczynnik	*Błąd stand.*	*z*	*wartość p*	
const	−2,92799	0,746180	−3,924	<0,0001	***
tryb	−0,304450	0,169854	−1,792	0,0731	*
srednia	0,492461	0,200156	2,460	0,0139	**
stypendium	0,502971	0,141746	3,548	0,0004	***
wiek	0,0161220	0,00959630	1,680	0,0930	*

Średn.aryt.zm.zależnej	0,375104	Odch.stand.zm.zależnej	0,484351
McFadden R-kwadrat	0,330201	Skorygowany R-kwadrat	0,223930
Logarytm wiarygodności	−773,1697	Kryt. inform. Akaike'a	1556,339
Kryt. bayes. Schwarza	1581,811	Kryt. Hannana-Quinna	1565,932

Źródło: opracowanie i obliczenia własne przy wykorzystaniu aplikacji Gretl.

Model regresji dla kierunku *Zarządzanie* przyjmuje zatem postać:

$$P(Y = 1/X_1, X_2, X_3, X_4) = \frac{1}{1 + e^{-(-2{,}92799 - 0{,}304450 \cdot X_1 + 0{,}492461 \cdot X_2 + 0{,}502971 \cdot X_3 + 0{,}0161220 \cdot X_4)}} \quad (5)$$

gdzie: X_1 – *tryb*, X_2 – *srednia*, $X_3 - stypendium, X_4 - wiek$

Weryfikacja zgodności modelu z danymi empirycznymi

Dla modeli binarnych stosuje się różne oceny stopnia dopasowania modelu do danych empirycznych. Najprostszą miarą jest współczynnik determinacji R^2 (skorygowany *R-kwadrat*, uwzględniający liczbę zmiennych w modelu). Można zastosować również współczynnik *R-*

kwadrat McFaddena. Należy zaznaczyć przy tym, że niski poziom wyjaśnienia zmienności jest cechą wszystkich modeli logitowych[11].

Wielu praktyków uważa, że o „dobroci" modelu decyduje trafność prognoz uzyskiwanych na jego podstawie[12]. Wykorzystuje się w tym celu miarę nazywaną *zliczeniowym R-kwadrat*. Wyznaczono liczbę przypadków poprawnej predykcji, która oznacza liczbę przypadków, dla których model dokonał poprawnej klasyfikacji, tzn. określił przynależność do klasy kodowanej jako 1 (kontynuacji studiów na poziomie SUM), gdy oszacowana wartość prawdopodobieństwa jest większa od wartości granicznej prawdopodobieństwa. Wartość „odcinającą" wyznaczono na poziomie 0,5. *Zliczeniowy R-kwadrat* określa udział poprawnie zaklasyfikowanych przypadków w łącznej liczbie przypadków. Podstawowym sposobem oceny modelu regresji logistycznej jest przedstawienie wyników klasyfikacji (prognozy) w formie tablicy trafień (tabela 6).

Tabela 6. Tablica trafień w modelu (próba ucząca)

Empiryczne (faktyczne)	Przewidywane		Razem
	0	1	
0	**524**	**229**	753
1	**229**	**223**	452
Razem	753	452	1205

Źródło: opracowanie i obliczenia własne przy wykorzystaniu aplikacji *Gretl*.

Zliczeniowy R-kwadrat w modelu wyniósł (524 + 223) / 1205 = 62,0%, co oznacza, że niemal 62% klasyfikacji przypadków okazała się być prawidłowa.

Trafność prognoz ocenić też można za pomocą *ilorazu szans*, obliczanego jako:

$(223 \cdot 524) / (229 \cdot 229) = 2{,}23.$

Należy zaznaczyć, że wartości ilorazu szans większe od 1 oznaczają, że prognozowanie na podstawie modelu jest lepsze od losowej (przypadkowej) klasyfikacji[13].

Możliwości prognostyczne modelu zweryfikowano posługując się danymi z roku 2018, w którym dyplom wyższego wykształcenia na poziomie licencjackim na kierunku *Zarządzanie* uzyskało 110 studentów (tabela 7).

[11] T. Kufel, *Ekonometria*…, op. cit., s. 146.
[12] P. Cichosz, *Data Mining Algorithms: Explained Using R*, Willey, Chichester 2015, s. 134-157.
[13] T. Kufel, *Ekonometria*…, op. cit., s. 146.

Tabela 7. Tablica trafień w modelu (próba weryfikująca)

Empiryczne (faktyczne)	Przewidywane		Razem
	0	1	
0	**28**	**28**	56
1	**21**	**67**	88
Razem	49	95	144

Źródło: opracowanie i obliczenia własne przy wykorzystaniu aplikacji MS Excel.

Trzeba zauważyć, że udział poprawnie zaklasyfikowanych przypadków w sumarycznej liczbie przypadków, czyli zliczeniowy R-kwadrat wzrósł do (28+67)/144 = 66%, zaś iloraz szans osiągnął wartość 3,19.

4. Interpretacja wyników oraz zastosowanie pozyskanej wiedzy do prognozowania liczby studentów kontynuujących kształcenie w macierzystej uczelni

Model regresji logistycznej dla kierunku *Zarządzanie* ujawnił 4 determinanty dalszego kształcenia na studiach II stopnia (przy przyjętym poziomie istotności 10%):

- *tryb* studiów (stacjonarne/niestacjonarne),
- średnia ocen ze studiów (*srednia*),
- pobieranie stypendium (*stypendium*),
- wiek absolwenta, w którym przystąpił do obrony pracy dyplomowej (*wiek*).

Na podstawie znaku parametru stojącego przy zmiennej (wzór 5) można określić kierunek wpływu tej zmiennej objaśniającej na prawdopodobieństwo kontynuacji studiów w murach macierzystej uczelni. Dodatni znak oszacowanego współczynnika regresji przy zmiennych *srednia*, *stypendium* oraz *wiek* oznacza, że czynnikami sprzyjającymi podjęciu studiów II stopnia są: osiąganie dobrych wyników w nauce, przyznanie stypendium oraz wzrost dojrzałości absolwenta. Wzrost średniej ocen o 1 stopień oznacza wzrost o 63,6% szansy kontynuacji kształcenia w uczelni (przy niezmiennym wpływie pozostałych zmiennych niezależnych). Uzyskanie stypendium oznacza wzrost szansy „sukcesu" o 65,4% (*ceteris paribus*). Dodatkowy rok w wieku absolwenta wiąże się ze wzrostem szansy pozostania na uczelni o 1,6% (zakładając, że pozostałe zmienne uwzględnione w modelu pozostają bez zmian). Ujemny znak współczynnika przy zmiennej tryb wskazuje, że przejście studenta na kształcenie w trybie niestacjonarnym zmniejsza prawdopodobieństwo przynależności tego absolwenta do klasy studiujących na poziomie magisterskim w uczelni (*ceteris paribus*). Szansa maleje o 26,2%.

Zbudowany model logistyczny zastosowano do przetworzonych danych studentów ostatniego semestru studiów I stopnia (których obrona pracy ustalona została na 2019 rok). Na podstawie oszacowanych parametrów modelu wyznaczono prawdopodobieństwa podjęcia przez nich studiów na poziomie magisterskim. Rysunek 2 przedstawia fragment arkusza obliczeniowego zastosowanego dla studentów badanego kierunku. Przyjęto wartość „odcinającą" na poziomie 0,5.

Rysunek 2. Fragment arkusza kalkulacyjnego, wykorzystanego do predykcji liczby studentów z kierunku *Zarządzanie* kontynuujących kształcenie na studiach magisterskich

tryb	*srednia*	*stypendium*	*wiek*	*logit*	*P*	*predykcja*
0	3,66	0	21	-0,78702074	0,312808732	0
1	4,29	1	47	0,14092269	0,535172484	1
0	3,99	0	21	-0,62450861	0,348756734	0
1	4,05	1	25	-0,33195195	0,41776576	0
1	4,16	0	26	-0,76463024	0,317641836	0
1	4,29	0	29	-0,65224431	0,342483966	0
1	4,74	1	25	0,00784614	0,501961525	1
1	4,63	1	41	0,21162743	0,55271028	1
1	4,08	0	29	-0,75566112	0,319589024	0
0	3,08	1	25	-0,50518912	0,376321983	0

Źródło: opracowanie i obliczenia własne.

Uzyskano wynik wskazujący, że co trzeci student ostatniego semestru studiów I stopnia badanego kierunku podejmie kształcenie na studiach II stopnia w macierzystej uczelni.

Podsumowanie

W opracowaniu podjęto próbę oszacowania liczby studentów uzupełniających studiów magisterskich w uczelni niepublicznej, rekrutujących się z grona jej absolwentów studiów licencjackich. Wykorzystano w tym celu dane gromadzone w systemach informatycznych uczelni i dokonano ich eksploracji z wykorzystaniem modelu regresji logistycznej. Zidentyfikowane zostały 4 czynniki, które mogą wpływać na decyzje studentów odnośnie kontynuacji studiów na poziomie SUM w macierzystej uczelni. Interpretacja i ocena parametrów modelu prowadzi do wniosku, że wysoka średnia ocen studentów, pozyskanie stypendium oraz dojrzałość absolwenta, zwiększają prawdopodobieństwo kontynuacji kształcenia w uczelni macierzystej. Szansę podjęcia przez absolwenta dalszych studiów zmniejsza kształcenie w trybie niestacjonarnym.

Przybliżono liczbę studentów, którzy zasilą szeregi przyszłych magistrów. Dla osób zarządzających uczelnią oszacowana wielkość pozwolić może na planowanie przyszłych działań o charakterze praktycznym.

Choć opracowany model wykazał swą przydatność, to jednak należy zwrócić uwagę na słabości predykcyjne modelu, wynikające między innymi z niewystarczającego, zakresu dostępnej informacji, a co za tym idzie niewielkiej liczby analizowanych zmiennych. Autorzy opracowania zdają sobie sprawę, że zaprezentowany model wymaga udoskonalenia. Wskazane jest wzbogacenie zbioru zmiennych tak, aby lepiej odwzorowywały modelowaną rzeczywistość. Wydaje się, że w szczególności brak jest w modelu informacji o sytuacji finansowej studenta, ważnej w przypadku podejmowania nauki na płatnych studiach w uczelni niepublicznej. Gromadzone w informatycznych systemach uczelnianych dane pozwoliły jedynie na niedoskonałe przybliżenie kondycji finansowej za pomocą informacji o przyznanych różnorodnych stypendiach. Ważny również wydaje się pomiar satysfakcji studenta z jakości świadczonych mu przez uczelnię usług. Równie istotną, choć trudną do pozyskania zmienną są informacje o dalszych losach absolwentów, którzy nie podjęli dalszego kształcenia w uczelni. Istotna wydaje się również informacja o dostępnych kierunkach studiowania w konkurencyjnych uczelniach. Należy również wspomnieć, że począwszy od roku 2020, wraz z wystąpieniem ogólnoświatowej pandemii, nastąpiła wymuszona zmiana sposobu prowadzenia zajęć dydaktycznych na uczelniach. W związku z tym dotychczasowe modele wymagają uwzględnienia nowych uwarunkowań i zmiennych. Wydaje się to być bardzo trudnym zadaniem, aczkolwiek określającym dalszy kierunek badań.

Literatura

[1] Cichosz P., *Data Mining Algorithms: Explained Using R*, Willey, Chichester 2015.
[2] Gawryś I., Trippner P., *Analiza poziomu rentowności przedsiębiorstwa na przykładzie niepublicznej uczelni wyższej w roku akademickim 2015/2016*, „Annales Universitatis Mariae Curie-Skłodowska", 2017, nr LI, 5.
[3] James G., Witten D., Hastie T., Tibshira R., *An Introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics, Springer Science+Business Media, New York 2013.
[4] Judge G. G., Hill C., Griffiths W. E., Lütkepohl H., Lee T., *The Theory and Practice of Econometrics*, John Wiley&Sons, New York 1985.
[5] Kufel T., *Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL*, Wydawnictwo Naukowe PWN, Warszawa 2011.
[6] Morzy T., *Eksploracja danych*, Wydawnictwo Naukowe PWN, Warszawa 2013.
[7] Nandeshwar A., Chaudhari S., *Enrollment Prediction Models Using Data Mining* http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf.
[8] Stanisz A., *Modele regresji logistycznej,* wyd. StatSoft Polska, Kraków 2016.
[9] Szeliga M., *Data Science i uczenie maszynowe*, Wydawnictwo Naukowe PWN, Warszawa 2017.
[10] Zieliński G., Lewandowski K., *Determinanty percepcji jakości usług edukacyjnych w perspektywie grup interesariuszy*, http://zif.wzr.pl/pim/2012_3_3_4.pdf.

Streszczenie

W artykule podjęto próbę wykorzystania modelu regresji logistycznej do oszacowania liczby absolwentów studiów I stopnia, którzy podejmą dalsze kształcenie na studiach magisterskich w macierzystej uczelni.

Dokonano eksploracji danych dostępnych w systemach informatycznych uczelni i na tej podstawie dokonano identyfikacji czynników mogących mieć wpływ na decyzje absolwentów odnośnie kontynuacji studiów na poziomie uzupełniających studiów magisterskich. Interpretacja i ocena parametrów zbudowanego modelu logitowego pozwoliła na sformułowanie wniosków o charakterze aplikacyjnym.

Słowa kluczowe

Logit, eksploracja danych, rekrutacja na uczelnię, regresja logistyczna, podejmowanie decyzji, przedsiębiorstwo.