STUDIA OECONOMICA POSNANIENSIA 2017, vol. 5, no. 12

DOI: 10.18559/SOEP.2017.12.3

Anna Lamek, Włodzimierz Lewoniewski
Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Informatyki Ekonomicznej

Autor do korespondencji: Włodzimierz Lewoniewski, wlodzimierz.lewoniewski@ue.poznan.pl

ZASTOSOWANIE REGRESJI LOGISTYCZNEJ W OCENIE JAKOŚCI INFORMACJI NA PRZYKŁADZIE WIKIPEDII

Streszczenie: Wykorzystanie regresji logistycznej w ocenie jakości danych może mieć szczególne znaczenie w dobie *big data*, gdzie mamy do czynienia z wieloma zmiennymi opisującymi dane zjawiska lub zachowania. Obliczenie rzeczywistej wartości informacji pozwala na wyeliminowanie tych zmiennych, które niewiele „wnoszą" do opisywanego zjawiska. Dzięki temu możliwa jest redukcja szumu informacyjnego i jednocześnie skupienie się na tych zmiennych, które najlepiej charakteryzują interesujące nas zjawisko, co także może przyczynić się do podejmowania właściwych decyzji. Wysoka jakość zmiennych sprzyja również budowaniu modeli prognostycznych, które pozwalają przewidzieć, jak konkretne dane będą wpływały na kształtowanie się zjawiska. W artykule zaprezentowano wykorzystanie regresji logistycznej w ocenie zmiennych opisujących jakość artykułów umieszczanych w Wikipedii w wersji angielskiej. Dokonano klasyfikacji zmiennych ze względu na uzyskany wskaźnik wartości informacyjnej (*IV – Information Value*) oraz dokonano oceny zdolności predykcyjnych. Przeprowadzone badanie może stanowić punkt wyjścia do porównania wyników z różnych wersji językowych Wikipedii.

Słowa kluczowe: *big data*, zarządzanie danymi, jakość informacji, regresja logistyczna, Wikipedia.

Klasyfikacja JEL: C8, D8, D82, L86.

APPLICATION LOGISTIC REGRESSION IN ASSESSING THE QUALITY OF INFORMATION – WIKIPEDIA ARTICLES CASE

Abstract: The use of the logistic regression in the assessment of the quality of data may have a significant impact on data management in the era of *big data*, where we are all dealing with a number of variables and amount of information describing some interesting phenomenon or behaviour. The calculation of actual an information value (IV) indicator allows to eliminate these variables which are irrelevant or just constitute an information overload. The article presents the use of logistic regression in the assessment of variables describing the quality of articles published on the English version of Wikipedia. A classification of variables because of the results of the information value indicator have been presented. Also the predictive capabilities of variables have been evaluated.

Keywords: big data, enterprise, data management, information value, logistic regression.

Wstęp

Wikipedia jest jednym z najpopularniejszych źródeł wiedzy na świecie. Obecnie ta ogólnodostępna encyklopedia jest na piątym miejscu w rankingu najczęściej odwiedzanych stron internetowych[1]. Artykuły w Wikipedii mogą być tworzone i edytowane przez dowolną osobę bez konieczności potwierdzenia swoich kompetencji w określonych obszarach. Ta encyklopedia nie posiada również centralnej redakcji, która mogłaby sprawdzać każde zmiany wprowadzone przez użytkowników (w tym anonimowych). Zasady wolnej encyklopedii coraz częściej są przedmiotem krytyki, szczególnie za zamieszczanie treści niskiej jakości. Niemniej jednak strony Wikipedii o znanych osobach, wydarzeniach, miejscach, firmach, produktach często pojawiają się jako pierwsze w wynikach wyszukiwania Google, Bing, Yandex i innych popularnych serwisów. Można się spodziewać, że jednocześnie czytelnicy Wikipedii i jej twórcy są zainteresowani wysoką jakością treści w niej zawartej. W roli twórców mogą występować znane firmy, które muszą dbać o obiektywne przedstawienie informacji o swoich produktach i usługach.

Regresja logistyczna pozwala ocenić wskaźnik wartości informacyjnej przy użyciu danych numerycznych i ich kategoryzacji. W statystycznej analizie danych czasami niezbędne jest, by określić, które z zestawu zmiennych są

[1] http://www.alexa.com/siteinfo/wikipedia.org.

„najlepsze" (tj. posiadają odpowiednią pojemność informacyjną), aby najlepiej opisać określone zachowanie czy zjawisko. Na przykład chcemy zidentyfikować, który z publikowanych w Wikipedii artykułów jest dobrej jakości, a który nie spełnia tego wymogu, jednocześnie wiedząc, że istnieje wiele zmiennych opisujących daną publikację. Często charakterystyki te niosą ze sobą nadmiar informacji (*information overload*), który trudno (również ze względu na binarną postać samej oceny – dobra/zła jakość publikacji) zidentyfikować i wyeliminować tradycyjnymi metodami. Konieczne jest, aby określić, które atrybuty mogą potencjalnie taki artykuł dobrze identyfikować, i dokonać odpowiedniej klasyfikacji. Celem artykułu jest zaprezentowanie wykorzystania regresji logistycznej w wyznaczeniu czynnika wartości informacji (*IV/Info-Value – Information Value*). Metoda ta przydaje się szczególnie w przypadku dużej liczby zmiennych niezależnych, a więc w momencie gdy ze względu na możliwą liczbę kombinacji trudne jest wyliczenie pojemności informacyjnej ogólnie znanymi sposobami, np. metodą Hellwiga.

1. Zmienne opisujące jakość artykułu w Wikipedii

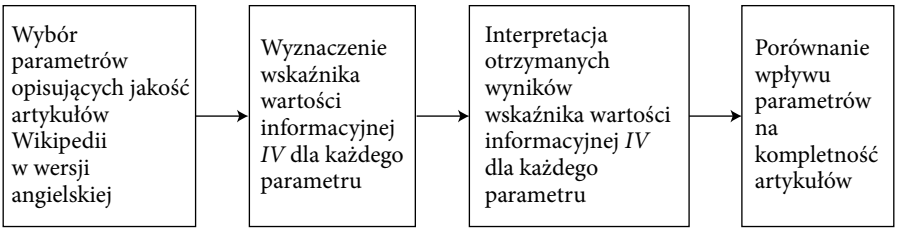
W różnych wersjach językowych Wikipedii dla artykułów najwyższej jakości istnieje odrębne wyróżnienie: w angielskiej wersji taki artykuł ma nazwę „Featured Article" (FA), w polskiej wersji odpowiednikiem jest nazwa „Artykuł na medal" (ANM). Istnieje również drugie wyróżnienie dla artykułów, które nie spełniają wszystkich kryteriów FA, ale zbliżają się do ich jakości: nazwa angielska to „Good Article" (GA), polska – „Dobry artykuł" (DA). Wyróżnienie FA lub GA w każdej wersji językowej artykuł może dostać dzięki pozytywnemu wynikowi głosowania, które zazwyczaj trwa około miesiąca. Jednak takich wyróżnionych artykułów jest bardzo mało – średnio w każdej wersji językowej udział ich wynosi około 0,07% całkowitej liczby artykułów.

W niektórych wersjach językowych artykuły oprócz wyróżnienia FA i GA mogą otrzymać inne (niższe) oceny, które mogą wskazywać na „dojrzałość" tych artykułów. Bez względu na to, że każda wersja językowa może mieć swój system klasyfikacji jakości artykułów, wszystkie stosują co najmniej dwie klasy dla artykułów najwyższej jakości – odpowiedniki FA i GA. Jednak duża część artykułów nie jest w ogóle oceniona, np. w polskiej edycji udział artykułów nieocenionych stanowi ponad 99% całkowitej ich liczby.

Wzrost popularności Wikipedii przyczynił się również do pojawienia się większej liczby prac naukowych dotyczących problematyki automatycznej oceny jakości artykułów w tej encyklopedii. Porównując różne parametry

artykułów wyróżnionych, można wnioskować o jakości innych (nieocenionych) artykułów. Na postawie literatury [Lih 2004; Stvilia i in. 2005; Hu i in. 2007; Wilkinson i Huberman 2007; Blumenstock 2008a; Blumenstock 2008b; Dalip i in. 2009; Lipka i Stein 2010; Dalip i in. 2011; Anderka 2013; Warncke-Wang, Cosley i Riedl 2013] oraz własnych badań [Lewoniewski, Węcel i Abramowicz 2015; Węcel i Lewoniewski 2015; Lewoniewski, Węcel i Abramowicz 2016] wybraliśmy sześć parametrów, które mogą wpływać na jakość artykułu i które stanowiły przedmiot badań:

- *referencje* – liczba wszystkich referencji, które są używane w treści artykułu;
- *odsłony* – liczba odsłon danego artykułu przez ostatnie 90 dni od momentu ekstrakcji danych;
- *liczba_obserwujących* – liczba użytkowników Wikipedii, którzy są bezzwłocznie informowani o wszelkich zmianach wprowadzonych do danego artykułu;
- *obrazki* – liczba obrazków, umieszczonych w artykule;
- *liczba_edycji* – liczba edycji artykułu od momentu powstania;
- pochodną złożoną *referencje/liczba_liter* – stosunek wszystkich referencji, które są używane w treści artykułu, do zmiennej *liczba_liter*, a więc liczby znaków, która jest używana w kodzie źródłowym.



Rysunek 1. Wizualizacja procedury badawczej

Podobnie jak w innych badaniach [Xu i Luo 2011; Lex i in. 2012; Warncke-Wang, Cosley i Riedl 2013; Lewoniewski, Węcel i Abramowicz 2015; Węcel i Lewoniewski 2015; Lewoniewski, Węcel i Abramowicz 2016] do budowania modelu będzie stosowana binarna zmienna objaśniana i jakość będzie modelowana jako prawdopodobieństwo przynależności do jednej z dwóch kategorii:

- 1 – kompletne artykuły: klasy FA i GA,
- 0 – niekompletne artykuły: wszystkie inne – rozwijające się (które należy dopracować) oraz nieocenione artykuły.

Wikipedia posiada serwis API, który zapewnia wygodny dostęp do danych i metadanych do artykułów za pomocą protokołu HTTP, za pośrednictwem

adresu URL, w różnych formatach (w tym XML, JSON). Serwis ten działa dla każdej wersji językowej i jest dostępny pod adresem określonym według szablonu: https://{język}.wikipedia.org/w/api.php?action={ustawienia}, gdzie {język} oznacza skrót wersji językowej, {ustawienia} – ustawienia zapytania[2].

W celu równoważenia analizowanych danych dobór liczebności próby trenującej odbywał się z zachowaniem rozkładu jednorodnego w taki sposób, aby pierwsza połowa liczebności kompletnych składała się z artykułów klasy FA. Drugą połowę stanowią artykuły z klasy GA. Zazwyczaj we wszystkich wersjach językowych Wikipedii klasa FA zawsze zawiera mniej artykułów niż GA. W przypadku angielskiej Wikipedii liczba artykułów klasy FA wynosi 4744 i jest mniejsza od liczby artykułów z klasy GA, zatem do próby trenującej losowo dobraliśmy 4744 artykuły z klasy GA. W ten sposób grupa kompletnych liczy 9488 artykułów w próbie trenującej. Liczebność grupy niekompletnych dobiera się na podstawie liczebności kompletnych. W wyniku równoważenia dla wersji angielskiej liczba artykułów w próbie trenującej wyniosła 18 976 artykułów (w tym 50% próby to artykuły kompletne, a reszta – niekompletne). Do próby testującej zostały dobrane wszystkie artykuły z klasą FA i GA (należące do kompletnych). Na podstawie liczebności tych artykułów została dobrana podobna liczba artykułów z innych (niższych) klas jakości (należących do niekompletnych). W rezultacie próba testująca zawierała około 58 tysięcy artykułów.

2. Wartość informacyjna

W licznych pracach [m.in. Brotherton i Lund 2013; Finlay 2010; Mays i Lynas 2011; Siddiqi 2006] opisujących modele scoringowe dotyczące np. zdolności kredytowej (gdzie również zmienna objaśniana ma postać binarną – przyznanie bądź nie kredytu) do oceny badanych cech wykorzystano współczynnik wartości informacyjnej (*IV – Information Value*), bazujący na regresji logistycznej. Wartość informacji zawartych w zmiennych objaśniających w celu opisu zmiennej objaśnianej odzwierciedla jakość tych zmiennych niezależnych, jak również ich moc prognostyczną.

Wynik wskaźnika WOE (*Weight of Evidence*), stanowiący element *IV*, pozwala tą metodą ocenić dobór każdego atrybutu badanej cechy zgodnie ze wzorem:

[2] Wszystkie możliwe ustawienia serwisu API można znaleźć na specjalnej stronie: https://pl.wikipedia.org/wiki/Specjalna:ApiSandbox.

$$WOE_i = \ln\left(\frac{g_i}{b_i}\right)\cdot 100. \tag{1}$$

Z kolei sama wartość informacyjna *IV* bada dodatkowo siłę predykcyjną zmiennej, co może wskazywać zarówno na jej przydatność, jak i możliwość tworzenia precyzyjnych modeli. Można ją wyznaczyć, stosując wyrażenie:

$$IV = \sum_{i=1}^{n}\left(g_i - b_i\right)\cdot \ln\left(\frac{g_i}{b_i}\right), \tag{2}$$

gdzie za przykładem Wikipedii interpretacja poszczególnych elementów jest następująca:

b_i – liczba artykułów ocenionych jako niekompletne (złe) przy uwzględnieniu konkretnej wartości lub przedziału wartości zmiennej objaśniającej X_i w stosunku do łącznej liczby niekompletnych artykułów w zbiorze całkowitym X (np. ile razy artykuły zawierające określoną liczbę (lub przedział) obrazków (zmienna *obrazki*) zostały ocenione jako niekompletne w stosunku do łącznej liczby niekompletnych artykułów),

g_i – liczba artykułów oceniona jako kompletne (dobre) przy uwzględnieniu konkretnej wartości lub przedziału wartości zmiennej objaśniającej X_i w stosunku do łącznej liczby niekompletnych artykułów w zbiorze całkowitym X.

W przypadku gdy $(g_i - b_i) = 0$ lub $\ln\left(\frac{g_i}{b_i}\right) = 0$ to jednostkowa wartość informacyjna przyjmuje wartość zero (nie ma znaczenia dla wartości sumarycznej).

Z reguły w modelach dąży się do tego, by brać pod uwagę zmienne o wysokim współczynniku *IV*, zaznaczając jednocześnie, aby ten współczynnik nie był zbyt wysoki, bo takie zmienne mogą zdominować model, co wiąże się z ryzykiem osłabienia stabilności modelu oraz pogorszenia jego precyzji. Zwolennicy wykorzystania regresji logistycznej w ocenie zmiennych stosują dwa podejścia. Pierwsze polega na wyborze tych zmiennych, które charakteryzują się średnią i silną zdolnością predykcyjną (przedziały wartości *IV* dla poszczególnych kategorii zostały przedstawione w tabeli 1. W drugim podejściu za „wartościowe" dla zaawansowanego budowania modeli uznawane są tylko te zmienne, które są predyktorami o średnim współczynniku *IV*. Przyjmuje się, że są to zmienne, które najlepiej charakteryzują dane zjawisko i nie wykazują niepożądanych cech dominacji w modelu.

Tabela 1. Interpretacja wartości współczynnika *Information Value IV*

Wartość *Information Value*	**Charakter zmiennej niezależnej, moc predykcyjna**
< 0,02	bezużyteczny predyktor
0,02–0,1	słaby predyktor
0,1–0,3	średni predyktor
0,3–0,5	mocny predyktor
> 0,5	czasem traktowany jako nadal mocny predyktor albo podejrzany – „zbyt dobry", aby mógł być wiarygodny, zmienna o cechach dominujących w modelu

Źródło: na podstawie: [Finlay 2010; Mays i Lynas 2011; Siddiqi 2006].

3. Ocena wartości informacji na przykładzie zmiennych opisujących artykuły Wikipedii

Do zaprezentowania metody oceny wartości informacyjnej wybrano zestaw danych opisujących jakość artykułów Wikipedii w wersji angielskiej. Zmienną objaśnianą jest ocena, czy dany artykuł jest kompletny, czy nie, inaczej mówiąc, czy jest oceniany jako artykuł dobry, czy zły (binarna postać recenzji). Obliczenia dotyczą kilku wybranych zmiennych, które zdaniem autorów mogą znacząco się różnić, jeśli chodzi o swoją wartość informacyjną i rolę zmiennej w budowaniu modeli predykcyjnych. Starano się, aby zmienne były reprezentantami różnych typów, zarówno tych pierwotnych, jak i złożonych, będącymi pochodną oryginałów. Pod uwagę wzięto takie zmienne, jak *referencje*, *liczba_obserwujących*, *liczba_edycji*, *odsłony*, *obrazki* oraz zmienną wtórną *referencja/liczba liter*. Sama metoda zostanie omówiona na przykładzie zmiennej *obrazki*.

Stworzenie tabeli przestawnej pozwoliło określić, ile razy w zbiorze danych pojawia się w zależności od liczby obrazków ocena niekompletny (0) lub kompletny (1) artykuł.

Tabela 2. Liczba kompletnych i niekompletnych artykułów w angielskiej wersji Wikipedii w zależności od liczby obrazków

Obrazki	**Niekompletny (0)**	**Kompletny (1)**	**Suma końcowa**
0	144	0	144
1	269	8	277
2	203	32	235
3	119	62	181

cd. tabeli 2

Obrazki	Niekompletny (0)	Kompletny (1)	Suma końcowa
4	85	69	154
5	53	69	122
6	25	71	96
7	16	64	80
8	19	59	78
9	9	56	65
10	10	55	65
11	9	61	70
12	4	60	64
13	5	32	37
14	3	25	28
15	4	41	45
16	2	22	24
17	1	13	14
18	1	25	26
19	1	14	15
20	2	10	12
21	1	11	12
22	2	10	12
23	0	7	7
24	2	8	10
25	0	8	8
26	3	6	9
27	1	12	13
28	1	4	5
29	0	9	9
30	0	10	10
31	0	8	8
32	0	4	4
33	1	7	8
34	1	4	5
35	0	4	4
36	0	6	6
37	0	3	3
38	0	1	1
39	0	1	1
40	0	3	3
41	0	3	3
42	0	1	1
43	0	2	2

cd. tabeli 2

Obrazki	Niekompletny (0)	Kompletny (1)	Suma końcowa
45	0	2	2
46	0	1	1
48	0	1	1
51	0	1	1
52	1	0	1
53	0	2	2
55	0	2	2
56	0	1	1
57	1	0	1
58	0	2	2
60	0	1	1
62	0	3	3
65	0	1	1
67	1	0	1
74	1	0	1
75	0	1	1
76	0	1	1
87	0	1	1
Suma końcowa	**1000**	**1000**	**2000**

Wskazane jest także obliczenie tzw. wskaźnika odpowiedzi (RR – *Responsive Rate*), który pokazuje, jaką część wszystkich wyników stanowią pozytywnie ocenione artykuły (szukamy przede wszystkim dobrych publikacji). Następnie dokonaliśmy obliczeń, jaki procent w każdej kategorii artykułów (% wśród kompletnych oraz % wśród niekompletnych) stanowią publikacje z określoną liczbą obrazków, a także jaką część populacji stanowi każdy z tych przypadków. Kolejnym krokiem była ocena wartości informacji w przypadku zmiennej *obrazki*, gdzie przy użyciu wzoru (2) uzyskano następujące wyniki.

Tabela 3. Wartość informacyjna w przypadku zmiennej *obrazki* – wyniki obliczeń

Obrazki	0	1	Suma końcowa	RR	Odsetek z „0”	Odsetek z „1”	Odsetek populacji	Wartość informacyjna
0	144	0	144	0	14,40	0,00	7,20	–
1	269	8	277	3	26,90	0,80	13,85	0,917485
2	203	32	235	14	20,30	3,20	11,75	0,315917
3	119	62	181	34	11,90	6,20	9,05	0,037163

cd. tabeli 3

Obrazki	0	1	Suma końcowa	RR	Odsetek z „0”	Odsetek z „1”	Odsetek populacji	Wartość informacyjna
4	85	69	154	45	8,50	6,90	7,70	0,003337
5	53	69	122	57	5,30	6,90	6,10	0,004221
6	25	71	96	74	2,50	7,10	4,80	0,048015
7	16	64	80	80	1,60	6,40	4,00	0,066542
8	19	59	78	76	1,90	5,90	3,90	0,045324
9	9	56	65	86	0,90	5,60	3,25	0,085922
10	10	55	65	85	1,00	5,50	3,25	0,076714
11	9	61	70	87	0,90	6,10	3,50	0,099510
12	4	60	64	94	0,40	6,00	3,20	0,151651
13	5	32	37	86	0,50	3,20	1,85	0,050120
14	3	25	28	89	0,30	2,50	1,40	0,046646
15	4	41	45	91	0,40	4,10	2,25	0,086109
16	2	22	24	92	0,20	2,20	1,20	0,047958
17	1	13	14	93	0,10	1,30	0,70	0,030779
18	1	25	26	96	0,10	2,50	1,30	0,077253
19	1	14	15	93	0,10	1,40	0,75	0,034308
20	2	10	12	83	0,20	1,00	0,60	0,012876
21	1	11	12	92	0,10	1,10	0,60	0,023979
22	2	10	12	83	0,20	1,00	0,60	0,012876
23	0	7	7	100	0,00	0,70	0,35	
24	2	8	10	80	0,20	0,80	0,50	0,008318
25	0	8	8	100	0,00	0,80	0,40	
26	3	6	9	67	0,30	0,60	0,45	0,002079
27	1	12	13	92	0,10	1,20	0,65	0,027334
28	1	4	5	80	0,10	0,40	0,25	0,004159
29	0	9	9	100	0,00	0,90	0,45	
30	0	10	10	100	0,00	1,00	0,50	
31	0	8	8	100	0,00	0,80	0,40	
32	0	4	4	100	0,00	0,40	0,20	
33	1	7	8	88	0,10	0,70	0,40	0,011675
34	1	4	5	80	0,10	0,40	0,25	0,004159
35	0	4	4	100	0,00	0,40	0,20	

cd. tabeli 3

Obrazki	0	1	Suma końcowa	RR	Odsetek z „0”	Odsetek z „1”	Odsetek populacji	Wartość informacyjna
36	0	6	6	100	0,00	0,60	0,30	
37	0	3	3	100	0,00	0,30	0,15	
38	0	1	1	100	0,00	0,10	0,05	
39	0	1	1	100	0,00	0,10	0,05	
40	0	3	3	100	0,00	0,30	0,15	
41	0	3	3	100	0,00	0,30	0,15	
42	0	1	1	100	0,00	0,10	0,05	
43	0	2	2	100	0,00	0,20	0,10	
45	0	2	2	100	0,00	0,20	0,10	
46	0	1	1	100	0,00	0,10	0,05	
48	0	1	1	100	0,00	0,10	0,05	
51	0	1	1	100	0,00	0,10	0,05	
52	1	0	1	0	0,10	0,00	0,05	
53	0	2	2	100	0,00	0,20	0,10	
55	0	2	2	100	0,00	0,20	0,10	
56	0	1	1	100	0,00	0,10	0,05	
57	1	0	1	0	0,10	0,00	0,05	
58	0	2	2	100	0,00	0,20	0,10	
60	0	1	1	100	0,00	0,10	0,05	
62	0	3	3	100	0,00	0,30	0,15	
65	0	1	1	100	0,00	0,10	0,05	
67	1	0	1	0	0,10	0,00	0,05	
74	1	0	1	0	0,10	0,00	0,05	
75	0	1	1	100	0,00	0,10	0,05	
76	0	1	1	100	0,00	0,10	0,05	
87	0	1	1	100	0,00	0,10	0,05	
Suma końcowa	**1000**	**1000**	**2000**	50				**2,33**

Okazuje się, że w przypadku zmiennej *obrazki* wartość informacyjna wynosi 2,33. W celu porównania wyników podobne obliczenia przeprowadzono dla pozostałych wybranych zmiennych, w tym jednej zmiennej złożonej. W tabeli 4 zaprezentowano uzyskane wyniki.

Tabela 4. Wartość informacyjna w przypadku różnych zmiennych objaśniających – zestawienie wyników

Zmienna	*IV* – wartość informacyjna
obrazki	2,33
referencje	2,95
liczba_edycji	0,68
liczba_obserwujących	0,28
odsłony	0,05
referencje/liczba_liter	1,69

4. Interpretacja wyników

W świetle przedstawionych wyników należy podkreślić, iż użycie regresji logistycznej pozwala na szczegółową ocenę zmiennych objaśniających pod kątem wartości informacyjnej, a w konsekwencji ich znaczenia w modelu. Obliczenia dotyczące przykładowych zmiennych na podstawie wzoru (2) pozwoliły wyznaczyć dla każdej z nich wartość wskaźnika informacyjnego. Okazuje się, że dwie zmienne – *referencje* oraz *obrazki*, a także zmienna złożona (również oparta na referencjach) są zaskakująco „dobre", innymi słowy, ich „moc" predykcyjna jest bardzo wysoka. Może to budzić pewne wątpliwości, czy aby zmienne te nie są zbyt silne i nie wykazują tendencji do dominacji w modelu. Stosunkowo mniejsze „podejrzenia" tej samej natury mogą się pojawić w stosunku do zmiennej *liczba_edycji*. W publikacjach dotyczących wartości informacyjnej pojawia się tu jednak pewna rozbieżność – w niektórych z nich (m.in. Brotherton i Lund 2013) zakłada się bowiem, że przekroczenie wartości 0,3 powoduje zaliczenie zmiennej do kategorii mocnych predyktorów, świadczących także o tym, że jest to cecha dobrze opisująca badane zjawisko i wcale nie oznacza dominacji danej cechy. Co do pozostałych przedziałów istnieje pełna zgodność, zatem zmienną *liczba_obserwujących* możemy zaliczyć do średnich, a *odsłony* do słabych predyktorów, jeśli chodzi o wpływ na jakość artykułu w Wikipedii.

Podsumowanie

Wykorzystanie regresji logistycznej w ocenie jakości danych może mieć szczególne znaczenie w dobie *big data*, gdzie mamy do czynienia z wieloma zmiennymi opisującymi dane zjawiska lub zachowania (wektor zmiennych

objaśniających na wejściu), które mają wpływ na podejmowanie decyzji. Wyjściowa ocena, o której decyduje duża liczba danych, ma z kolei często charakter binarny (dobra/zła jakość, kupić/nie kupić dany towar, udzielić kredytu czy nie). Obliczenie rzeczywistej wartości informacji pozwala na wyeliminowanie tych zmiennych, które niewiele wnoszą do opisywanego zjawiska. Dzięki temu możliwa jest redukcja szumu informacyjnego i jednocześnie skupienie się na tych zmiennych, które najlepiej charakteryzują interesujące nas zdarzenie i w dużej mierze przyczyniają się do podejmowania właściwych decyzji. Wysoka jakość zmiennych sprzyja także budowaniu modeli prognostycznych, pozwalających przewidzieć, jak konkretne dane będą wpływały na kształtowanie się zjawiska, co również stanowi solidny fundament wspierający dokonywanie wyborów. Interesujące może się okazać wykorzystanie tej metody w momencie grupowania danych w przedziały wartości, co może znacząco zmienić globalną wartość informacyjną. Ciekawe może być też porównanie jej efektywności z innymi sposobami oceny jakości artykułów, jak również wyznaczenie oceny, która decyduje o tym, że dana publikacja stanowi już dobre źródło informacji, chociaż czasami nie spełnia całkowicie kryteriów odpowiadających ocenie „najlepszy". Kluczowe może się również okazać, jak dobór odpowiednich zmiennych przy użyciu funkcji logistycznej może wpływać na precyzję modeli. Wspomniane zagadnienia będą stanowiły przedmiot dalszych badań.

Bibliografia

Anderka, M., 2013, *Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia*, PhD. Bauhaus-Universität, Weimar.

Belanger, D., Betser, J., 2013, *Architecting the Enterprise via Big Data Analytics*, in: Liebowitz, J. (ed.), *Big Data and Business Analytics*, CRC Press, Taylor & Francis Group, Boca Raton, s. 1–20.

Berry, D., 2012, *Unstructured Data: Challenge or Asset?*, http://www.zdnet.com/article/unstructured-data-challenge-or-asset/ [dostęp: kwiecień 2016].

Blumenstock, J.E., 2008a, *Automatically Assessing the Quality of Wikipedia Articles*, School of Information, UC Berkeley.

Blumenstock, J.E., 2008b, *Size Matters: Word Count as a Measure of Quality on Wikipedia*, w: *Proceedings of the 17th International Conference on World Wide Web*, s. 1095–1096.

Brotherton, D., Lund, B., 2013, *Information Value Statistic*, MidWest SAS® Users Group conference materials, Paper AA-14-2013.

Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P., 2009, *Automatic Quality Assessment of Content Created Collaboratively by Web Communities: A Case Study of Wikipedia*, in: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, s. 295–304.

Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P., 2011, *Automatic Assessment of Document Quality in Web Collaborative Digital Libraries*, Journal of Data and Information Quality, vol. 2, no. 3, s. 1–30.

Finlay, S., 2010, *Credit Scoring, Response Modelling and Insurance Rating*, Palgrave MacMillan, New York.

Hu, M., Lim, E.-P., Sun, A., Lauw, H.W., Vuong, B.-Q., 2007, *Measuring Article Quality in Wikipedia*, in: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, s. 243–252.

Lewoniewski, W., Węcel, K., Abramowicz, W., 2015, *Analiza porównawcza modeli jakości informacji w narodowych wersjach Wikipedii*, w: Porębska-Miąc, T. (red.), *Systemy Wspomagania Organizacji SWO*, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice, s. 133–154.

Lewoniewski, W., Węcel, K., Abramowicz, W., 2016, *Quality and Importance of Wikipedia Articles in Different Languages*, in: Dregvaite, G., Damasevicius, R. (eds.), *Information and Software Technologies. ICIST 2016*, Communications in Computer and Information Science, iss. 639, s. 613-624.

Lex, E., Voelske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., Stein, B., Granitzer, M., 2012, *Measuring the Quality of Web Content Using Factual Information*, in: *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web*, s. 7.

Lih, A., 2004, *Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Resource*, in: *5th International Symposium on Online Journalism*, s. 31.

Lipka, N., Stein, B., 2010, *Identifying Featured Articles in Wikipedia: Writing Style Matters*, in: *Proceedings of the 19th International Conference on World Wide Web*, s. 1147–1148.

Mays, E., Lynas, N., 2011, *Credit Scoring for Risk Managers: The Handbook for Lenders*, 2nd ed., South Western, Thomson.

Siddiqi, N., 2006, *Credit Risk Scorecards*, John Wiley & Sons, Hoboken, NJ.

Stvilia, B., Twidale, M.B., Smith, L.C., Gasser, L., 2005, *Assessing Information Quality of a Community-based Encyclopedia*, in: *Proceedings of the Internetional Conference on Information Quality*, s. 442–454.

Warncke-Wang, M., Cosley, D., Riedl, J., 2013, *Tell Me More: An Actionable Quality Model for Wikipedia*, in: *Proceeding of the 9th International Symposium on Open Collaboration*, s. 1–10.

Węcel, K., Lewoniewski, W., 2015, *Modelling the Quality of Attributes in Wikipedia Infoboxes*, in: Abramowicz, W. (ed.), *Business Information Systems Workshops. BIS 2015*, Lecture Notes in Business Information Processing, iss. 228, s. 308-320.

Wilkinson, D.M., Huberman, B.A., 2007, *Cooperation and Quality in Wikipedia*, in: *Proceedings of the 2007 International Symposium on Wikis*, s. 157–164.

Xu, Y., Luo, T., 2011, *Measuring Article Quality in Wikipedia: Lexical Clue Model*, in: *IEEE Symposium on Web Society 19*, s. 141–146.