Analýza návštevníkov spravodajskej webstránky s využitím logistickej regresie

Romana Šipoldová[1], Erik Šoltés[2]

Abstrakt
Článok analyzuje návštevníkov webovej stránky so spravodajstvom, ktorí by sa s najväčšou pravdepodobnosťou mohli stať predplatiteľmi plateného obsahu na danej web stránke. Za účelom predikovania návštevníkov je vytvorený model binárnej logistickej regresie za použitia štatistického softvéru SAS. Cieľom článku je identifikovanie faktorov, ktoré signifikantne ovplyvňujú pravdepodobnosť predplatenia si plateného obsahu na spravodajskej webstránke. Ďalším cieľom je následné vytvorenie profilu užívateľa, ktorý sa stane predplatiteľom plateného obsahu (zamknutých článkov na spravodajskej webstránke) a rovnako tak profil užívateľa, pri ktorom je pravdepodobnosť predplatenia si tohto obsahu veľmi nízka.

Kľúčové slová
digitálny marketing, logistická regresia, spravodajská webstránka

Abstract
The article analyses news website visitors who are most likely to become subscribers to paid content on that website. In order to predict the visitors, a binary logistic regression model is created using SAS statistical software. The aim of the article is to identify the factors that significantly affect the probability of subscribing to paid content on the news website. Another goal is the subsequent creation of a user profile that becomes a subscriber to paid content (locked articles on the news website), as well as a user profile for who the probability of subscribing to this content is very low.

Key words
digital marketing, logistic regression, news website

JEL classification
C2, M3

1 Úvod

Hlavnými piliermi úspešnej digitálnej marketingovej kampane sú dáta, analytika, personalizácia, optimalizácia a automatizácia. Digitálny marketing je v súčasnosti nielen pre veľké firmy a značky, ale v porovnaní s tradičným marketingom je dostupný aj menším podnikom, a to efektívne a za prijateľnú cenu. Veľkou výhodou digitálneho marketingu je poskytovanie personalizovaného obsahu až na úroveň jednotlivca. Podľa prieskumu od spoločnosti Infogroup (Zawacki, 2019), ktorý bol zameraný na súčasné postoje a preferencie spotrebiteľov týkajúcich sa personalizovanej reklamy, je personalizovaná reklama dôležitou až očakávanou súčasťou reklamnej komunikácie. Spoločnosti si uvedomujú, aká dôležitá súčasť reklamnej komunikácie je personalizovaná reklama, ale len málokto dokáže pri komunikácii s existujúcimi alebo potenciálnymi zákazníkmi zabezpečiť personalizovanosť a relevantnosť tejto reklamy pre dané publikum. Pomocou digitálneho marketingu je možné zasiahnuť cieľovú

[1] Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra štatistiky, Dolnozemská cesta 1, 852 35 Bratislava, romana.sipoldova@euba.sk.
[2] Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra štatistiky, Dolnozemská cesta 1, 852 35 Bratislava, erik.soltes@euba.sk.

skupinu merateľným a cenovo efektívnym spôsobom, ale len za predpokladu optimalizovanej marketingovej stratégie.

Dalessandro a kol. (2015) sa vo svojej práci venujú skutočnosti, že online reklama a jej optimalizácia je v porovnaní s optimalizáciou cez tradičné médiá oveľa prístupnejšia. V práci vytvorili model logistickej regresie na predikciu nákupov na internetovej stránke na základe návštevnosti stránky, nie na základe prekliku cez reklamný banner. Podľa záverov, ktoré z práce získali, je takýto model oveľa presnejší pri predikcii nákupov ako model vychádzajúci len z kliknutí na daný reklamný banner. Aj Constantin (2015) a Miralles-Pechuán a kol. (2018) vo svojej práci využívajú model logistickej regresie, a to v marketingovom výskume pri tvorbe marketingovej stratégie a pri optimalizácii online reklamných kampaní. Olson a Chae (2012) využili vo svojej práci, zameranej na segmentáciu zákazníkov, model logistickej regresie a model rozhodovacieho stromu. Predmetom výskumu mnohých prác boli aj rôzne typy cielení využité v online reklamnej kampani. Chen a Stallaert (2014), podobne ako my, využili vo svojom výskume behaviorálne cielenie a okrem toho analyzovali aj ekonomické dôsledky tohto cielenia na zadávateľov reklamy a poskytovateľov reklamného priestoru. De Bock a Van den Poel (2010) vo svojej práci využili demografické cielenie, pričom predikovali profily používateľov internetovej stránky pomocou modelu náhodného lesa.

V článku sa venujeme využitiu metódy strojového učenia na vytvorenie prediktívneho modelu, ktorý bude slúžiť pri odhadovaní toho, či internetový užívateľ patrí do cieľovej skupiny, ktorá je charakterizovaná určitými behaviorálnymi znakmi a na ktorú má byť cielená online reklama. Konkrétne sa zameriavame na predikciu predplatiteľov plateného (zamknutého) obsahu na webovej stránke so spravodajstvom. Vstupné údaje boli poskytnuté marketingovou agentúrou GroupM a obsahujú informácie o 3808 návštevníkoch webovej stránky. Táto stránka neponúka žiadny obsah ani články zadarmo, tzn. že ak chce návštevník čítať nejaký článok, musí sa najskôr zaregistrovať a získať buď dočasný prístup k vybraným článkom (tzv. trial verzia), alebo si predplatí prístup k článkom na určité obdobie. Dočasný prístup je obmedzený iba na jeden mesiac a po mesiaci si buď čitateľ predplatí prístup k obsahu spravodajskej stránky (zvyčajne na obdobie jedného roka), alebo stratí prístup k obsahu článkov. Rozhodnutie o predplatení prístupu k článkom je motivované obsahom alebo typom článkov, ktoré návštevníka zaujímajú. Môžu však na to pôsobiť aj iné faktory. Dáta, ktoré máme k dispozícii, poskytujú behaviorálne informácie o návštevníkoch stránky. Cieľom článku je predikovať, či si návštevník stránky predplatí platený obsah spravodajskej stránky alebo nie.

2 Príprava dát

Vstupné premenné (tab. 1) je pred vstupom do analýzy potrebné upraviť. Z premenných *Hour_of_day* a *Day_of_week* vytvoríme nové premenné, ktoré budú opisovať afinitu[3] cieľovej skupiny voči hodine počas dňa a dňa v týždni. Tzn., že vytvoríme premennú *Affinity_HoD* a *Affinity_DoW*, ktoré budú numerické.

Následne sa zameriame na URL (*Uniform Resource Locator*) konkrétnych stránok a referenčných stránok. Referenčné stránky nás informujú o tom, z akého zdroja na danú web stránku užívateľ prišiel. Z pôvodnej premennej *Http_referer* sme vytvorili dve premenné – najskôr prvú premennú *Source_type*, ktorá hovorí o tom, či užívateľ prišiel z webovej stránky alebo z aplikácie. Druhá premenná bude *Source_platform*, ktorá bude detailnejšie opisovať platformu, z ktorej užívateľ prišiel.

[3] Affinita (Index affinity) – predstavuje podiel sledovanej vlastnosti u cieľovej skupiny a u celkovej populácie. Ak je index affinity väčší ako 1 (väčší ako 100 %), sledovaná vlastnosť sa vyskytuje viac u cieľovej skupiny ako u celkovej populácie.

Tab. 1: Zoznam vstupných premenných v dátovom súbore

Názov premennej	Opis premennej
USER_ID	Unikátne ID užívateľa
BROWSER	ID prehliadača
DEVICE_TYPE	Typ využívaného zariadenia
GEO_DMA	Geo región
OPERATING_SYSTEM	ID verzie operačného systému
HOUR_OF_DAY	Hodina návštevy web stránky
DAY_OF_WEEK	Deň v týždni návštevy web stránky
URL	Prvá stránka s článkom, ktorú navštívil užívateľ stránky so spravodajstvom
HTTP_REFERER	Referenčná stránka pre prvú navštívenú stránku spravodajskej webstránky
AUTOMOTIVE, BOOK_AND_LITERATURE, BUSINESS_AND_FINANCIAL, ...	Premenné klasifikujúce obsah navštívenej stránky pomocou taxonómie IAB[4] (spolu 30 premenných)
TARGET	Cieľová premenná – či si užívateľ predplatil platený obsah webstránky (1) alebo nie (0)

Zdroj: GroupM data source; Vlastné spracovanie v SAS EM

Každá URL adresa je zložená z niekoľkých častí. Opíšeme ich na príklade:

https://www.thetimes.co.uk/article/paradise-lost-the-latest-from-the- ... - hf6mvjhps

ktorý môžeme vo všeobecnosti zapísať takto:

https://www.thetimes.co.uk/ [sekcia_1/ sekcia_2] / [článok]

Názvy sekcie 1, resp. sekcie 2 budú predstavovať obmeny (kategórie) nových premenných – *Sekcia_1* a *Sekcia_1_a_2*. Zvyšná časť premennej *URL*, ktorú sme označili ako *článok*, bude predmetom analýzy textu, tzv. Text Miningu. Vzhľadom na rozsiahlosť analýzy túto časť v článku neuvádzame (Šipoldová, 2022).

Z Text Mining analýzy a následnej zhlukovej analýzy sme získali 5 zhlukov, ktoré predstavujú 5 nových binárnych premenných, každú za jeden zhluk (*Cluster_**). Tieto budú ďalej využité ako vstupné premenné pri vytváraní prediktívneho modelu.

3 Model logistickej regresie

Prediktívny model, ktorý sme vytvorili, je model logistickej regresie. Keďže závislá premenná bola binárna a hovorila o tom, či si užívateľ predplatí platený obsah (*Target* = 1) alebo nie (*Target* = 0), použili sme model binárnej logistickej regresie. Väčšina vstupných premenných, vstupujúcich do modelu, sú kategoriálne premenné, ale do modelu vstupujú aj numerické premenné *Affinity_HoD* a *Affinity_DoW*. Na vytvorenie prediktívneho modelu budeme využívať softvér SAS Enterprise Miner.

Vysvetľujúce premenné, ktoré boli binárne, mali dve obmeny, a to obmenu 1 pre užívateľov, ktorí patrili do príslušnej kategórie, ktorú vystihuje názov binárnej premennej a obmenu 0 pre tých užívateľov, ktorí nepatrili do tejto kategórie. Ako referenčná kategória bola nastavená kategória s hodnotou 0.

U ostatných premenných, ktoré boli kategoriálne, boli nastavené referenčné kategórie nasledovne:

- premenná *Browser* mala referenčnú kategóriu *Samsung Browser*,
- premenná *Device_type* mala referenčnú kategóriu *3 – Tablet*,

[4] IAB Taxonómia predstavuje taxonómiu obsahu, ktorú možno použiť pri opise obsahu danej webovej stránky. Bližší opis možno nájsť na stránke IAB: https://iabtechlab.com/standards/content-taxonomy/.

- premenná *Geo_DMA* mala referenčnú kategóriu *1*,
- premenná *Operating_system* mala referenčnú kategóriu *Catalina 10.15*,
- premenná *Sekcia_1* mala referenčnú kategóriu *article*,
- premenná *Sekcia_1_a_2* mala referenčnú kategóriu *article_*,
- premenná *Source_type* mala referenčne kategóriu *Web* a
- premenná *Source_platform* mala referenčnú kategóriu *Google/Googlesearch*.

Najskôr rozdelíme vstupný súbor na dve časti - trénovaciu množinu a validačnú množinu v pomere 70:30, pomocou stratifikovaného náhodného výberu. Tento pomer sa zvyčajne používa pri metódach učenia sa s učiteľom (Xu, Goodacre, 2018). Pri vytváraní modelu sme vybrali ako metódu výberu premenných krokovú regresiu (*stepwise selection*) a ako kritérium na výber nezávislej premennej (*selection criterion*) Profit/Loss.

Vo výsledkoch najskôr dostávame časť, ktorá overuje štatistickú významnosť modelu ako celku (tab. 2) a štatistickú významnosť vplyvu jednotlivých premenných (tab. 3).

Tab. 2: Test významnosti modelu ako celku

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood		Likelihood		
Intercept Only	Intercept & Covariates	Ratio Chi-Square	DF	Pr > ChiSq
3680.645	2797.354	883.2902	47	<.0001

Zdroj: GroupM data source; Vlastné spracovanie v SAS EM

Na hladine významnosti $\alpha = 0{,}05$ bol model ako celok štatisticky významný, čo znamená, že sme zamietli nulovú hypotézu a prijali alternatívnu hypotézu. Keďže sme využili metódu krokovej regresie na selekciu premenných, tak niektoré premenné boli z modelu vylúčené a zostali len premenné, ktoré majú štatisticky významný vplyv.

Tab. 3: Test významnosti vplyvu jednotlivých premenných

Effect	DF	Wald Chi-Square	Pr > ChiSq
Affinity_DoW	1	6.8063	0.0091
Affinity_HoD	1	8.0180	0.0046
Books_and_Literature	1	8.2120	0.0042
Browser	4	17.0987	0.0018
CLUSTER_2	0	.	.
DeviceType	2	49.0223	<.0001
Education	1	38.8753	<.0001
GeoDMA	14	30.9495	0.0056
Healthy_Living	1	19.0288	<.0001
Medical_Health	1	8.2852	0.0040
Source_platform	15	93.0004	<.0001
Source_type	1	161.7613	<.0001
Style___Fashion	1	12.5383	0.0004
sekcia 1 a 2	0	.	.

Zdroj: GroupM data source; Vlastné spracovanie v SAS EM

Ďalej sme vo výstupe dostali odhady parametrov modelu binárnej logistickej regresie a odhady zodpovedajúcich pomerov šancí (Tab. 4).

Tab. 4: Odhad parametrov a pomerov šancí

Parameter		Beta	p-value	Odds Ratio
Intercept		-2.621	<.0001	–
Affinity_HoD		1.226	0.0046	3.407
Affinity_DoW		0.874	0.0091	2.395
Browser	Edge vs Samsung Br.	0.990	0.0499	2.691
Browser	Chrome vs Samsung Br.	0.796	0.0002	2.217
Browser	Safari vs Samsung Br.	0.649	0.0044	1.913
Browser	Unknown vs Samsung Br.	0.138	0.7483	1.148
CLUSTER_2	1 vs 0	-0.245	0.0433	0.783
DeviceType[5]	1 vs 3	1.284	<.0001	3.612
DeviceType	2 vs 3	0.474	0.0453	1.606
GeoDMA	2658 vs 1	-0.183	0.6611	0.832
GeoDMA	2669 vs 1	-0.529	0.0164	0.589
GeoDMA	2670 vs 1	-0.630	0.0129	0.532
GeoDMA	2665 vs 1	-0.662	0.0116	0.516
GeoDMA	2672 vs 1	-0.693	0.0154	0.500
GeoDMA	2671 vs 1	-0.756	0.0475	0.470
GeoDMA	2666 vs 1	-0.812	0.0179	0.444
GeoDMA	2660 vs 1	-0.870	0.0138	0.419
GeoDMA	2659 vs 1	-0.907	0.0013	0.404
GeoDMA	2668 vs 1	-0.969	0.0005	0.379
GeoDMA	2664 vs 1	-0.984	0.0001	0.374
GeoDMA	2667 vs 1	-0.987	<.0001	0.373
GeoDMA	2663 vs 1	-1.009	0.0043	0.365
GeoDMA	2662 vs 1	-1.650	0.0379	0.192
Source_platform	gmail vs G/GS[6]	3.638	0.0011	37.998
Source_platform	apple vs G/GS	1.012	0.0856	2.751
Source_platform	linkedin vs G/GS	0.940	0.0988	2.561
Source_platform	internal vs G/GS	0.775	0.0991	2.170
Source_platform	twitter vs G/GS	0.556	0.0045	1.743
Source_platform	ecosia vs G/GS	0.447	0.6053	1.563
Source_platform	instagram vs G/GS	0.327	0.7138	1.387
Source_platform	theguardian vs G/GS	0.325	0.7520	1.384
Source_platform	yahoo vs G/GS	-0.118	0.8737	0.889
Source_platform	undefined vs G/GS	-0.172	0.2160	0.842
Source_platform	bing vs G/GS	-0.41	0.4600	0.664
Source_platform	others vs G/GS	-0.681	0.0302	0.506
Source_platform	facebook vs G/GS	-1.121	<.0001	0.326
Source_platform	googlearticlesforyou vs G/GS	-1.269	<.0001	0.281
Source_platform	googlenews vs G/GS	-3.274	0.0019	0.038
Source_type	App vs Web	-1.838	<.0001	0.159
sekcia_1_a_2	1_Home_ vs 4_article_	1.664	<.0001	5.280
sekcia_1_a_2	3_others vs 4_article_	1.502	0.0002	4.492
sekcia_1_a_2	2_edition_ vs 4_article_	0.099	0.6086	1.104
Education	1 vs 0	1.922	<.0001	6.834
Healthy_Living	1 vs 0	1.091	<.0001	2.977
Style_and_Fashion	1 vs 0	0.924	0.0004	2.519
Books_and_Literature	1 vs 0	0.692	0.0042	1.997
Medical_Health	1 vs 0	-0.765	0.0040	0.465

Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v SAS EM

Kategoriálne premenné, ktoré vstupovali do modelu, boli nahradené umelými premennými, ktorých počet je vždy o 1 nižší, ako je počet kategórií danej premennej. Keďže

[5] DeviceType obmeny: 1 = dektop a laptop, 2 = mobilný telefón, 3 = tablet
[6] G/GS = Google/Googlesearch

sme mali veľký počet premenných, na interpretáciu sme využili len niektoré vybrané pomery šancí, ktoré boli interpretované za podmienky ceteris paribus (c. p.) – všetky ostatné vysvetľujúce premenné, ktoré boli do modelu zaradené, zostávajú na konštantnej úrovni.

Podiel pravdepodobnosti, že si užívateľ predplatí platený obsah na spravodajskej webstránke a pravdepodobnosti, že si ho nepredplatí je šanca. Táto šanca je pri užívateľoch, ktorí použili na prezeranie článku desktop alebo laptop, 3,612-násobne vyššia ako u tých, ktorí využili tablet. Uvedená šanca je u užívateľa, ktorý navštevuje články s tematikou Vzdelávania 6,834-násobne vyššia, ako u užívateľa, ktorý tieto články nenavštevuje a 6,289-násobne vyššia (1/0,159) u užívateľa, ktorý navštívil článok prostredníctvom webového prehliadača oproti návšteve cez aplikáciu.

Premenná *Source_platform* hovorí o tom, z akej platformy užívateľ prišiel. Ako referenčnú kategóriu sme zvolili vyhľadávač Google. Najskôr sme uviedli tie šance, kde bola daná platforma „úspešnejšia“ ako vyhľadávač Google, čiže všetky interpretované šance sú porovnávané so šancou prislúchajúcou k užívateľovi, ktorý sa preklikol na daný článok cez vyhľadávač Google. Pravdepodobnosť, že si užívateľ predplatí platený obsah, oproti pravdepodobnosti, že si ho nepredplatí, je:

- 37,998-násobne vyššia, ak užívateľ prišiel cez emailového klienta od spoločnosti Google – cez Gmail,
- 2,751-násobne vyššia, ak užívateľ prišiel zo zariadenia od spoločnosti Apple,
- 2,561-násobne vyššia, ak užívateľ prišiel cez príspevok alebo reklamu na sociálnej sieti LinkedIn,
- 2,170-násobne vyššia, ak užívateľ prišiel priamo cez spravodajskú webstránku thetimes alebo cez iný článok na danej webstránke,
- 1,743-násobne vyššia, ak užívateľ prišiel cez príspevok alebo reklamu na sociálnej sieti Twitter,

oproti tomu, ak používateľ prišiel cez vyhľadávač Google. Naopak, pravdepodobnosť predplatenia plateného obsahu bola pri prekliku cez platformu Google:

- 26,316-násobne vyššia oproti tomu, ak používateľ prišiel cez správy od spoločnosti Google (Google news),
- 3,067-násobne vyššia v porovnaní s tým, ak používateľ prišiel cez sociálnu sieť facebook,
- 1,976-násobne vyššia v porovnaní s tým, ak používateľ prišiel cez inú webovú stránku (napr. reddit, bbc.co.uk a pod.).

Pomery šancí pri ostatných platformách boli štatisticky nevýznamné.

Pravdepodobnosť, že si používateľ predplatí platený obsah na spravodajskej webstránke, oproti pravdepodobnosti, že si ho nepredplatí, je v závislosti od zaradenia článku do sekcie oproti sekcii article_ nasledujúca:

- 5,280-násobne vyššia, ak je článok zaradený do sekcie Home_ a
- 4,492-násobne vyššia, ak je článok zaradený do sekcie others.

Z premenných, ktoré charakterizujú klasifikáciu stránok podľa IAB, malo štatisticky významný prínos práve päť klasifikačných premenných. Šanca, že si používateľ predplatí platený obsah, je:

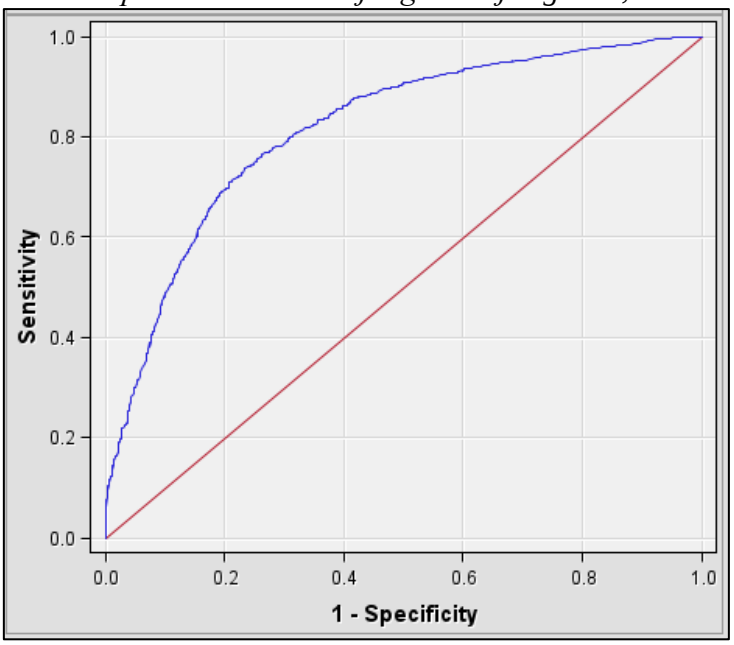
- 6,834-násobne vyššia, ak čítaný článok patril do kategórie Vzdelanie (*Education*), ako keď do tejto kategórie nepatril,
- 2,977-násobne vyššia, ak patril článok do kategórie Zdravý životný štýl (*Healthy Living*), ako keď do tejto kategórie nepatril,
- 2,519-násobne vyššia, ak patril článok do kategórie Štýl a móda (*Style and Fashion*), ako keď do tejto kategórie nepatril,
- 1,997-násobne vyššia, ak patril do kategórie Knihy a literatúra (*Books and Literature*), ako keď do tejto kategórie nepatril,

- o 53,47% nižšia, ak patril do kategórie Lekárstvo/Zdravie (*Medical_Health*), ako keď do tejto kategórie nepatril.

Z Text Mining analýzy vyšiel štatisticky významný iba zhluk č. 2, ktorý spájal články s tematikou pandémie covid-19. Ostatné zhluky boli štatisticky nevýznamné. Pravdepodobnosť, že si užívateľ predplatí platený obsah oproti pravdepodobnosti, že si ho nepredplatí, je 1,277-násobne vyššia (1/0,783), ak nepatrí do zhluku č. 2, čiže ak čítaný článok nepatril medzi články s témou o covid-19.

Vo výstupe zo SAS EM sme dostali aj graf ROC krivky (Obr. 1). ROC krivka vyjadruje na osi *x* podiel nepredplatiteľov nesprávne zaradených do kategórie predplatiteľov a všetkých prípadov, ktoré patria do kategórie nepredplatiteľov. Na osi *y* sú hodnoty, ktoré vyjadrujú podiel správne zaradených predplatiteľov do tejto kategórie a všetkých užívateľov, ktorí sú predikovaní ako predplatitelia. Veľkosť ROC indexu pre model logistickej regresie má veľkosť 0,815.

Obr. 1: ROC krivka pre model binárnej logistickej regresie, ROC index=0,815



Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v SAS EM

Predikované hodnoty vysvetľovanej premennej vychádzajú z predikovanej podmienenej pravdepodobnosti vypočítanej podľa hodnôt vysvetľujúcich premenných (vstupujúcich do modelu) pri danom pozorovaní. Nech p_{ij} je pravdepodobnosť, že i-té pozorovanie patrí do j-tej kategórie (Allison, 2012). Všeobecný model logistickej regresie má potom tvar:

$$\ln\left(\frac{p_{ij}}{p_{iJ}}\right) = \boldsymbol{\beta}_j \mathbf{x}_i \qquad j = 1, 2, \ldots, (J-1) \tag{1}$$

kde $\mathbf{x}_i$ je stĺpcový vektor vysvetľujúcich premenných a $\boldsymbol{\beta}_j$ je riadkový vektor koeficientov j-tej kategórie. Každá kategória sa porovnáva s kategóriou J. Úpravou dostávame rovnicu:

$$p_{ij} = \frac{e^{\boldsymbol{\beta}_j \mathbf{x}_i}}{1 + \sum_{k=1}^{J-1} e^{\boldsymbol{\beta}_k \mathbf{x}_i}} \qquad j = 1,2, \ldots, (J-1) \tag{2}$$

Tab. 5: Hodnoty vysvetľujúcej premennej Target použité na výpočet predikovanej podmienenej pravdepodobnosti

Predikovaná podmienená pravdepodobnosť			
Parameter	**Kategórie**	**Pravdepodobnosť $\hat{p}$**	
		Užívateľ - typické hodnoty	**Užívateľ - netypické hodnoty**
Predikovaná podmienená pravdepodobnosť		**0.99999478**	**0.00011195**
Intercept		1	1
Affinity_DoW		1.21	0.83
Affinity_HoD		1.25	0.47
Books_and_Literature	**1**	1	0
Browser	**Unknown**	0	0
Browser	**Chrome**	0	0
Browser	**Safari**	0	0
Browser	**Edge**	1	0
CLUSTER_2	**1**	0	1
DeviceType	**1**	1	0
DeviceType	**2**	0	0
Education	**1**	1	0
GeoDMA	**1**	1	0
GeoDMA	**2658**	0	0
GeoDMA	**2659**	0	0
GeoDMA	**2660**	0	0
GeoDMA	**2662**	0	1
GeoDMA	**2663**	0	0
GeoDMA	**2664**	0	0
GeoDMA	**2665**	0	0
GeoDMA	**2666**	0	0
GeoDMA	**2667**	0	0
GeoDMA	**2668**	0	0
GeoDMA	**2669**	0	0
GeoDMA	**2670**	0	0
GeoDMA	**2671**	0	0
Healthy_Living	**1**	1	0
Medical_Health	**1**	0	1
Source_platform	**01_apple**	0	0
Source_platform	**02_bing**	0	0
Source_platform	**04_ecosia**	0	0
Source_platform	**05_facebook**	0	0
Source_platform	**06_gmail**	1	0
Source_platform	**07_googlearticlesforyou**	0	0
Source_platform	**08_googlenews**	0	1
Source_platform	**09_instagram**	0	0
Source_platform	**10_internal**	0	0
Source_platform	**11_linkedin**	0	0
Source_platform	**12_others**	0	0
Source_platform	**13_theguardian**	0	0
Source_platform	**14_twitter**	0	0
Source_platform	**15_undefined**	0	0
Source_platform	**16_yahoo**	0	0
Source_type	**App**	0	1
Style_and_Fashion	**1**	1	0
sekcia_1_a_2	**1_Home_**	1	0
sekcia_1_a_2	**2_edition_**	0	0
sekcia_1_a_2	**3_others_**	0	0

Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v Microsoft Excel

Podľa vzorca (2) s využitím odhadov parametrov logitového modelu uvedených v Tab. 5 sme vypočítali pravdepodobnosť, že si užívateľ predplatí platený obsah na spravodajskej webstránke. Na výpočet bol využitý vektor vysvetľujúcich premenných, kde sme najskôr využili typické hodnoty pre danú cieľovú skupinu a pri druhom výpočte sme využili hodnoty vysvetľujúcich premenných, ktoré nie sú typické pre cieľovú skupinu. Hodnoty vidíme v Tab. 5.

Pri výpočte s typickými hodnotami pre užívateľa bola pravdepodobnosť, že si predplatí platený obsah na spravodajskej webstránke veľmi vysoká, a to $\hat{p}$ = 0,99999478, teda 99,999478 %. Užívateľ s najvyššou pravdepodobnosťou predplatenia si plateného obsahu navštívil daný článok v utorok v doobedných hodinách (najmä okolo 11 hodiny) prostredníctvom desktopu alebo laptopu, kde využil web namiesto aplikácie, konkrétne prehliadač Microsoft Edge. Referenčným zdrojom bol emailový klient od Google – Gmail. Článok bol podľa IAB klasifikácie zaradený buď do kategórie Vzdelávanie, Knihy a literatúra, Zdravý životný štýl alebo Štýl a móda.

Naopak, pri výpočte podmienenej pravdepodobnosti s netypickými hodnotami pre užívateľa bola táto pravdepodobnosť $\hat{p}$ = 0,00011195, teda 0,011195 %. Takýto užívateľ navštívil článok v piatok okolo 3 hodiny ráno prostredníctvom tabletu a cez aplikáciu, kde využil prehliadač Samsung Browser. Na článok sa dostal prostredníctvom Google News, čo je stránka alebo aplikácia od spoločnosti Google, ktorá ponúka užívateľom články na čítanie. Článok bol podľa IAB klasifikácie zaradený do kategórie Lekárstvo alebo Zdravie.

Tab. 6: Profily užívateľov s najvyššou a najnižšou pravdepodobnosťou predplatenia si plateného obsahu

Premenná	Najvyššia pravdepodobnosť predplatenia	Najnižšia pravdepodobnosť predplatenia
Source_Type	Web	Web
Device_Type	Desktop alebo Laptop	Desktop alebo Laptop
Source_Platform	Gmail	Google/Googlesearch
Sekcia_1_a_2	Home_	article_
Browser	Edge	X
IAB klasifikácia	Vzdelávanie, Knihy a literatúra, Zdravý životný štýl, Štýl a móda	Vzdelávanie
Hour of Day	11	X
Day of Week	Utorok	X

Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v SAS EM

4 Záver

V dnešnom svete moderných technológií prebiehajú inovácie v každej oblasti. Dáta sú cenným a dôležitým zdrojom pre každú spoločnosť, hoci nie každá ich využíva efektívnym spôsobom. Osloviť spotrebiteľov v dnešnej široko konkurenčnej spoločnosti je čoraz náročnejšie. Tradičné médiá sú napriek svojej efektívnosti spojené s vysokými nákladmi, ktoré si väčšinou menšie firmy nemôžu dovoliť investovať. Preto sa do popredia dostávajú digitálne médiá a digitálny marketing, ktorý umožňuje prístup na konkurenčný trh za dostupné náklady veľmi efektívnym spôsobom. Základným pilierom úspešného digitálneho marketingu sú dáta a kvalitné prediktívne modely, ktoré dokážu poskytovať pomerne presnú personalizáciu až na úroveň jednotlivca.

Článok bol zameraný na vytvorenie prediktívneho modelu, ktorý bude čo najlepšie predikovať užívateľov spravodajskej webstránky, ktorí sa s najväčšou pravdepodobnosťou

stanú predplatiteľmi plateného obsahu. Vytvoreniu prediktívneho modelu binárnej logistickej regresie predchádzala úprava vstupných údajov, ktorá okrem vytvárania nových premenných zahŕňala aj analýzu textu – Text Mining. Na základe získaných výsledkov môžeme konštatovať, že model predikoval cieľovú skupinu veľmi dobre, čo dokazuje aj ROC index, ktorého veľkosť bola 0,815. Získané výsledky ďalej hovorili aj o najdôležitejších faktoroch pri predikovaní cieľovej skupiny, medzi ktoré patrili: *Source_type, Device_Type, Source_Platform, Sekcia_1_a_2, Browser* a *Education*. Pri predplatení si plateného článku na spravodajskej webstránke boli najdôležitejšie faktory tie, či prišiel užívateľ na danú webstránku z inej webovej stránky alebo z aplikácie (*Source_type*), aký typ zariadenia využíval – desktop alebo laptop, mobilný telefón alebo tablet (*Device_type*), z akej stránky užívateľ prišiel (*Source_platform*), v ktorý deň v týždni a hodinu daný článok čítal, do akej sekcie bol článok zaradený (*Sekcia_1_a_2*), aký webový prehliadač užívateľ využíval (*Browser*) a či bol článok zaradený, podľa IAB klasifikácie, do kategórie Vzdelávanie (*Education*).

Podľa uvedených faktorov užívateľ, ktorý sa s najväčšou pravdepodobnosťou stane predplatiteľom plateného obsahu na spravodajskej webstránke, si daný článok prezeral v utorok okolo 11 hodiny ráno. Pri čítaní využil webstránku namiesto aplikácie, desktop alebo laptop a prehliadač od spoločnosti Microsoft – Edge. Na stránku prišiel prostredníctvom mailového klienta od spoločnosti Google – Gmail. Článok bol zahrnutý do sekcie Home_ a témou bolo Vzdelávanie, Knihy a literatúra, Zdravý životný štýl a Štýl a móda.

Vytvorený prediktívny model môže ďalej slúžiť ako model pre programatický nákup pri online reklamnej kampani, kde bude slúžiť na oslovenie cieľovej skupiny efektívnym a cenovo prijateľným spôsobom.

Literatúra

[1] Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS institute.

[2] Chen, J., stallaert, J. (2014). An economic analysis of online advertising using behavioral targeting. *MIS Quarterly*, 38(2), 429-449.

[3] Constantin, C. (2015). Using the Logistic Regression model in supporting decisions of establishing marketing strategies. *Economic Sciences*, 8(2), 43-50.

[4] Dalessandro, D., et al. (2015). Evaluating and optimizing online advertising: Forget the click, but there are good proxies. *Big data*, 3(2), 90-102.

[5] De Bock, K., Van Den Poel, D. (2010). Predicting website audience demographics forweb advertising targeting using multi-website clickstream data. *Fundamenta Informaticae*, 98(1), 49-70.

[6] Miralles-Pechuán, L., ponce, H., Martínez-Villaseñor, L. (2018). A novel methodology for optimizing display advertising campaigns using genetic algorithms. *Electronic Commerce Research and Applications*, 27, 39-51.

[7] Olson, D. L., chae, B. (2012). Direct marketing decision support through predictive customer response modeling. *Decision Support Systems*, 54(1), 443-451.

[8] Šipoldová, R. (2022). Using Text Mining to Analyse Web Addresses (URLS). *Economic and Social Development: 77th International Scientific Conference on Economic and Social*, 71-80.

[9] Xu, Y., Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, 2(3), 249-262.

[10] Zawacki, T. (2019). *Why Consumers Prefer Personalization*. Webpage: https://multichannelmerchant.com/blog/why-consumers-prefer-personalization/. (accessed on 14.06.2020).