EKONOMICKÁ UNIVERZITA V BRATISLAVE

Fakulta hospodárskej informatiky

Evidenčné číslo: 103005/D/2021/36086129771119620

CIELENIE ONLINE REKLAMY S VYUŽITÍM ŠTATISTICKÝCH METÓD

Dizertačná práca

2022 **Ing. Romana Šipoldová**

EKONOMICKÁ UNIVERZITA V BRATISLAVE

Fakulta hospodárskej informatiky

CIELENIE ONLINE REKLAMY S VYUŽITÍM ŠTATISTICKÝCH METÓD

Dizertačná práca

Študijný program:	Kvantitatívne metódy v ekonómii
Študijný odbor:	Ekonómia a manažment
Školiace pracovisko:	Katedra štatistiky
Školiteľ záverečnej práce:	prof. Mgr. Erik Šoltés, PhD.

Bratislava 2022 **Ing. Romana Šipoldová**

Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Ing. Romana Šipoldová
Študijný program: kvantitatívne metódy v ekonómii (Jednoodborové štúdium, doktorandské III. st., denná forma)
Študijný odbor: ekonómia a manažment
Typ záverečnej práce: Dizertačná záverečná práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Cielenie online reklamy s využitím štatistických metód

Literatúra: Odporúčaná literatúra:

Allison, P. D. (2012). Logistic Regression using SAS. Theory and Application (2nd ed.). North Carolina, USA: SAS Institute.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 16(5), 412-424.

Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. European Journal of Operational Research, 195(1), 1-16.

Braun, M., & Moe, W. W. (2013). Online display advertising: Modeling the effects of multiple creatives and individual impression histories. Marketing Science, 32(5), 753-767.

Dalessandro, B., Hook, R., Perlich, C., & Provost, F. (2015). Evaluating and optimizing online advertising: Forget the click, but there are good proxies. Big data, 3(2), 90-102.

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.

Dave, K., & Varma, V. (2014). Computational advertising: Techniques for targeting relevant ads. Foundations and Trends® in Information Retrieval, 8(4–5), 263-418.

Chen, J., & Stallaert, J. (2014). An economic analysis of online advertising using behavioral targeting. MIS Quarterly, 38(2), 429-449.

Kim, K., & Timm, N. (2006). Univariate and multivariate general linear models: theory and applications with SAS. Chapman and Hall/CRC.

Powers, D.M.W., (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technology, 2(1), 37-63.

Ramon, C. L., Stroup, W. W., & Freund, R. J. (2010). SAS for Linear Models (4th Revised ed.). North Carolina, USA: SAS Institute.

Tharwat, A. (2018). Classification assessment methods. Applied Computing and Informatics. In press, corrected proof. https://doi.org/10.1016/j.aci.2018.08.003. Retrieved April 22, 2019, from https://www.sciencedirect.com/science/article/pii/S2210832718301546

Wooldridge, J. M. (2013). Introductory econometrics: A modern approach (5th ed.). Mason: South-Western.

Cieľ: Cieľom práce je vytvoriť a porovnať prediktívne modely, ktoré budú slúžiť na optimalizáciu online kampane a na efektívne oslovenie cieľovej skupiny.

Anotácia: Online reklama v súčasnosti využíva rôzny typy cielenia (demografické, behaviorálne, kontextuálne, geografické atď.), vďaka čomu môže byť takáto forma reklamy veľmi efektívna. Aj uvedená skutočnosť je príčinou toho, že digitálny marketing v dnešnej dobe zaznamenáva veľký rozmach. Na dosiahnutie čo najefektívnejšej online kampane je však potrebné na základe správania užívateľa internetu predikovať, že ide o osobu, ktorá spadá do cieľovej skupiny, na ktorú je zameraná reklamná kampaň. Dizertačná práca má poskytnúť sumarizáciu metodík prediktívnych modelov ako sú model logistickej regresie, model rozhodovacieho stromu a model „náhodného lesa“ (random forest) a ich aplikáciu v online marketingu. Očakáva sa, že modely budú založené na kvalitnej databáze, preto významnou časťou práce bude príprava dát. Kvalita modelov bude porovnaná prostredníctvom klasifikačných mier, a to buď binomických alebo multinomických v závislosti od charakteru cieľovej premennej v uvedených modeloch.

Školiteľ: prof. Mgr. Erik Šoltés, PhD.
Katedra: KŠ FHI - Katedra štatistiky FHI
Vedúci katedry: doc. Ing. Mária Vojtková, PhD.

Dátum zadania: 10.06.2019

Dátum schválenia: 12.06.2019 doc. RNDr. Mária Bilíková, PhD.
predseda odborovej komisie

Čestné vyhlásenie

Čestne vyhlasujem, že záverečnú prácu som vypracovala samostatne a že som uviedla všetku použitú literatúru.

Dátum: ....

Poďakovanie

V prvom rade by som sa rada poďakovala svojmu školiteľovi dizertačnej práce prof. Mgr. Erikovi Šoltésovi, PhD. za jeho čas, odborné usmernenia, ochotu a cenné rady pri písaní dizertačnej práce, ľudský a mimoriadne ústretový prístup a podporu, ktorú mi bol kedykoľvek ochotný poskytnúť. Ďalej by som chcela poďakovať svojej rodine a priateľom, ktorí ma neustále podporovali a vytvárali mi počas štúdia priaznivé prostredie. Poďakovanie patrí aj členom katedry štatistiky FHI EU v Bratislave, na čele s doc. Ing. Máriou Vojtkovou, PhD. za ich nápomocnosť a ústretovosť, ktorú mi počas štúdia poskytli. Na záver by som chcela poďakovať aj spoločnosti SAS Slovakia, s. r. o., ktorá mi poskytla softvér SAS pre účely jeho využitia pri tvorbe dizertačnej práce.

Ďakujem.

Abstrakt

Ing. ŠIPOLDOVÁ, Romana: *Cielenie online reklamy s využitím štatistických metód*. – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra štatistiky. – Školiteľ záverečnej práce: prof. Mgr. Erik Šoltés, PhD. – Bratislava: FHI EU, 2022, 119 s.

Cieľom dizertačnej práce je identifikovanie profilu internetových užívateľov, ktorí patria do cieľovej skupiny, ktorá je charakterizovaná určitými behaviorálnymi znakmi a na ktorú má byť cielená online reklama, pomocou prediktívneho modelovania. Cieľ je naplnený v niekoľkých etapách. V prvej etape sa oboznámime s oblasťou digitálneho marketingu z hľadiska základných pojmov a jeho výhod a nevýhod. Ďalej sa zameriame na rozdelenie online kampaní, z hľadiska nákupu online reklamy a z hľadiska cielenia online reklamy. V druhej etape zosumarizujeme tri vybrané metódy strojového učenia, konkrétne model binárnej logistickej regresie, model rozhodovacieho stromu (*decision tree*) a model náhodného lesa (*random forest*). Okrem toho v tejto časti opisujeme miery využívané na hodnotenie jednotlivých modelov a porovnávajúcich ich kvalitu. Tretia etapa spočíva v aplikácii metód strojového učenia na vytvorenie prediktívnych modelov a na následný výber toho najlepšieho prediktívneho modelu. Skladá sa z viacerých čiastkových cieľov. Prvým cieľom je príprava dát, ktoré budú slúžiť na vytvorenie modelov, čo zahŕňa bližšiu analýzu vstupných premenných a ich prípadnú úpravu. Druhým cieľom je vytvorenie troch prediktívnych modelov, ktoré opíšeme v druhej etape. Pomocou mier porovnávajúcich kvalitu modelu vyberieme ten, ktorý má najlepšiu prediktívnu schopnosť. Tretím čiastkovým cieľom je vytvorenie profilu predplatiteľa plateného obsahu, na základe vytvorených prediktívnych modelov.

Práca je rozdelená do 5 kapitol. Obsahuje 17 grafov, 10 tabuliek a 18 obrázkov. Prvá kapitola opisuje súčasný stav riešenej problematiky doma a v zahraničí. Druhá kapitola sa venuje cieľom práce. Tretia kapitola je zameraná na metodiku práce a metódy skúmania, ktoré využijeme pri aplikácii metód strojového učenia na vytvorenie prediktívnych modelov. Jednotlivé podkapitoly sa venujú modelu logistickej regresie, modelu rozhodovacieho stromu a modelu náhodného lesa. Štvrtá kapitola je venovaná praktickej aplikácii získaných poznatkov. V každej kapitole je opísaný postup vytvorenia daného prediktívneho modelu a ich následné porovnanie na základe kritérií

porovnávajúcich úspešnosť a kvalitu prediktívnych modelov. Posledná kapitola je venovaná diskusii ohľadne vytvorenia prediktívnych modelov a využitia metód strojového učenia pri aplikácii v oblasti marketingu. Výsledný model získaný aplikáciou metód strojového učenia na vytvorenie prediktívnych modelov, predstavuje prediktívny model, ktorý bude využitý v procese programatického nákupu na predikciu návštevníkov webovej stránky so spravodajstvom, na ktorých bude cielená online reklamná kampaň. Cieľom tejto reklamnej kampane bude osloviť takých návštevníkov, ktorí by sa s najväčšou pravdepodobnosťou mohli stať predplatiteľmi plateného obsahu na spravodajskej webstránke. Napriek tomu, že model nie je univerzálny, výsledky poskytujú návod na aplikáciu metód a postupov aj pri iných webových stránkach s následným využitím tohto modelu ako modelu pre programatický nákup pri online reklamných kampaniach.

Kľúčové slová:
online marketing, cieľová skupina, programatický nákup, logistická regresia, rozhodovacie stromy, náhodný les

Abstract

Ing. ŠIPOLDOVÁ, Romana: *Targeting of online advertising using statistical methods.* – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Statistics. – Thesis supervisor: prof. Mgr. Erik Šoltés, PhD. – Bratislava: FEI UE, 2022, 119 p.

The aim of the doctoral thesis is to identify the profile of Internet users who belong to the target group, which is characterized by certain behavioral features and to which online advertising should be targeted, using predictive modeling. The goal is fulfilled in several stages. In the first stage, we will get acquainted with the field of digital marketing in terms of basic concepts and its advantages and disadvantages. We will also focus on the division of online campaigns, in terms of purchase of online advertising and in terms of targeting online advertising. In the second stage, we summarize three selected machine learning methods, namely the binary logistic regression model, the decision tree model and the random forest model. In addition, in this section we describe the measures used to evaluate individual models and compare their quality. The third stage consists in the application of machine learning methods to create predictive models and the subsequent selection of the best predictive model. It consists of several partial goals. The first partial goal is to prepare data that will be used to create models, which includes a closer analysis of input variables and their possible adjustment. The second partial goal is to create three predictive models, which we will describe in the second stage. Using measures comparing the quality of the model, we select the one that has the best predictive ability. The third partial goal is to create a subscriber's profile for paid content, based on the created predictive models.

The thesis is divided into 5 chapters. It contains 17 graphs, 10 tables and 18 pictures. The first chapter describes the current state of the issue in our country and abroad. The second chapter deals with the goals of the work. The third chapter focuses on the work methodology and research methods, which we will use in the application of machine learning methods to create predictive models. The individual subchapters deal with the logistic regression model, the decision tree model and the random forest model. The fourth chapter deals with the practical application of the acquired knowledge. Each chapter describes the process of creating a predictive model and their subsequent

comparison based on statistics which measure the success and quality of predictive models. The last chapter deals with the discussion of the creation of predictive models and the use of machine learning methods for application in marketing area. The resulting model, obtained by applying machine learning methods to create predictive models, is a predictive model capable that will be used in the process of programmatic buying to predict visitors of the news website on which the online advertising campaign will be targeted. The goal of this ad campaign will be to reach those visitors who are most likely to be subscribers to paid content on the news website. Although the model is not universal, the results provide guidance on the application of methods and procedures to other websites, with the subsequent use of this model as a model for programmatic buying in online advertising campaigns.

Key words:
online marketing, target group, programmatic buying, logistic regression, decision trees, random forest

O B S A H str.

Úvod

Analytika, dáta, personalizácia, automatizácia a optimalizácia. Toto sú hlavné piliere úspechu digitálnych marketingových kampaní. Digitálny marketing je v dnešnej dobe neoddeliteľnou súčasťou veľkých a úspešných značiek. Nie je však určený len pre veľké firmy a 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. Zasiahnuť cieľovú skupinu je možné merateľným a cenovo efektívnym spôsobom, ale len za predpokladu optimalizovanej marketingovej stratégie. Zadávatelia reklamy a marketingové agentúry sa snažia svoje publikum osloviť čo najpresnejšie reklamou, ktorá je tomuto publiku adresovaná.

V našej práci sa najskôr venujeme základným pojmom, ktoré sa s digitálnym marketingom a reklamou spájajú a opisom hlavných výhod a nevýhod, ktoré so sebou digitálny marketing prináša. Je dôležité opísať aj spôsoby nákupu online kampaní, typy nákupu a typy cielenia online reklamy. Konkrétne sa budeme venovať programatickému nákupu reklamy, ktorý sa v zahraničí, ale aj na Slovensku stáva čoraz viac rozšíreným. Typy cielenia online reklamy sú rôzne a keďže cieľom je presnosť a nákladová efektivita, je dôležité poznať všetky možnosti cielenia reklamy, ktoré zadávatelia reklamy a reklamné agentúry majú. Okrem najrozšírenejšieho broad cielenia, ktoré nemá konkrétne obmedzenia, sa venujeme aj napr. behaviorálnemu, demografickému alebo tzv. Look alike modelingu. Niekedy sa používa aj kombinácia viacerých typov cielení. Aby bolo možné využiť a nastaviť cielenie čo najpresnejšie a najefektívnejšie, je potrebné mať k dispozícii aj dostatočné množstvo dát, ktoré poslúžia na vytvorenie a spresnenie prediktívnych modelov.

Hlavná časť práce sa venuje opisu metód strojového učenia (*machine learningu*) a následne vytvoreniu prediktívnych modelov, ktoré budú slúžiť na predikciu cieľovej skupiny, pri využití dát z internetovej stránky. Konkrétne sa zameriame na model logistickej regresie, model rozhodovacieho stromu (*decision tree*) a model náhodného lesa (*random forest*), ktoré patria medzi algoritmy strojového učenia. Pomocou vhodných mier na hodnotenie kvality modelu sa vyberie najlepší model, ktorý bude slúžiť na predikciu cieľovej skupiny. Cieľom je využiť vytvorený prediktívny model na cielenie online reklamy a oslovenie budúcou online reklamnou kampaňou ďalších potenciálnych zákazníkov čo najefektívnejšie a s čo najnižšími nákladmi na reklamnú kampaň.

1 Súčasný stav riešene problematiky doma a v zahraničí

Dnešných spotrebiteľov zaplavuje viac informácií a ponúk než kedykoľvek predtým. V silnom konkurenčnom prostredí je čoraz náročnejšie zaujať spotrebiteľov produktami a výrobkami. Preto sa v praxi a veľmi často aj v odbornej a vedeckej literatúre využívajú pokročilé kvantitatívne metódy, ktoré poskytujú sofistikované a kvalitné prediktívne modely na zefektívnenie online reklamy. Digitálny marketing poskytuje prístup na masový trh za výhodnejšiu cenu, v porovnaní s reklamou v tradičných médiách (televízia, tlač, outdoor[1]). Na základe analyticky založených možností segmentácie zákazníkov je možné urobiť marketingové kampane cielenejšie s vyššou mierou odozvy. Metódy strojového učenia sa implementujú v oblasti digitálneho marketingu po celom svete, pričom konkrétne to zahŕňa použitie dát, obsahu a online kanálov na zvýšenie produktivity a lepšie pochopenie cieľového publika.

Súčasťou digitálneho marketingu je online marketing, často nazývaný aj internetový marketing, ktorý je nielen celosvetovo, ale aj na Slovensku veľmi rozšírený. Expanziu internetového marketingu dokazujú aj nárasty nákladov na internetovú reklamu a nárast v počte internetovej populácie. V roku 2019 predstavovali výdavky do internetovej reklamy na Slovensku 136,83 mil. eur, čo je o 244 % viac v porovnaní s rokom 2012, kedy tvorili výdavky do internetovej reklamy 56,02 mil. eur. (IAB Slovakia, 2020). V roku 2019 bolo online 76 % Slovákov vo veku 12-79 a využívalo internet ako médium. V absolútnych číslach ide o 3,5 mil. Slovákov, čo v porovnaní s rokom 2012 predstavuje nárast o 22 % (IAB Slovakia, 2021).

Výdavky do reklamy sa nezvyšujú len v rámci Slovenska, celosvetovo môžeme pozorovať podobný trend. V roku 2018 boli výdavky do online reklamy približne 227 mld. amerických dolárov, čo prevýšilo investície do tradičnej televíznej reklamy o 21 %. Zatiaľ čo v roku 2000 tvorili investície do internetovej reklamy iba 7 % investícií do televíznej reklamy, v roku 2016 boli investície do obidvoch médií približne na rovnakej úrovni (okolo 180 mld. dolárov). Hoci výdavky do televíznej reklamy neustále rastú, tento trend je oveľa pomalší ako v prípade výdavkov do online reklamy (Šoltés, Táborecká-Petrovičová, Šipoldová, 2020).

Nárast internetovej reklamy je spojený aj s výhodami, ktoré prináša, ako merateľnosť, nižšie náklady, presnejšie a efektívnejšie možnosti cielenia, globálny

[1] Outdoor (out-of-home, OOH) – vonkajšia reklama

rozsah, nepretržitá dostupnosť a sledovanie výsledkov v reálnom čase. Hoci má online marketing aj svoje nedostatky, ako bannerová slepota, využívanie ad-blockov, ktoré zabraňujú zobrazovaniu online reklamy alebo vysoká konkurencia, v porovnaní s reklamou v tradičných médiách dokáže internetová reklamná kampaň zasiahnuť relevantné publikum za oveľa nižšie náklady, presnejšie a efektívnejšie. Skutočnosť, že online reklama a jej optimalizácia je oveľa prístupnejšia ako optimalizácia cez tradičné médiá, spomínajú vo svojej práci napríklad Dalessandro a kol. (2015), ktorí na optimalizáciu online reklamy využívajú logistickú regresiu. Vytvorili model na predikciu nákupov na internetovej stránke na základe návštevnosti stránky, nie na základe prekliku cez reklamný banner. Podľa záverov z práce 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. Ďalšia práca, ktorá sa zaoberá vplyvom data miningu, konkrétne RFM analýzy (*recency, frequency, monetary*), modelu logistickej regresie a modelu rozhodovacieho stromu, na segmentáciu v priamom marketingu, pochádza od McCarty a Hastak (2006). V ich štúdii sú použité dva dátové súbory – jeden obsahuje dáta od zásielkovej spoločnosti a druhý od neziskovej organizácie. Cieľom bola segmentácia v priamom marketingu, a teda vytvorenie takého prediktívneho modelu, ktorý predpovedá ľudí s najvyššou tendenciou si niečo kúpiť alebo v prípade neziskovej organizácie, prispieť tejto organizácii. Výsledkom štúdie bolo zistenie, že oba modely, model logistickej regresie aj rozhodovacieho stromu, dosahovali pri daných údajoch oveľa lepšie výsledky ako RFM analýza. Bose a Chen (2009), ktorí sa tiež venovali oblasti priameho marketingu, skúmali silné a slabé stránky modelov, ktoré boli založené na data miningu a modelov, ktoré boli založené na regresii. Oblasti priameho marketingu sa venovali aj Karim a Rahman (2013), ktorí sa však vo svojej práci sústredili na model rozhodovacieho stromu a na Bayesovský algoritmus. Skúmali, ako môže data mining pomôcť spoločnostiam pri napĺňaní marketingových cieľov. Olson a Chae (2012) tiež používali vo svojej práci, zameranej na segmentáciu zákazníkov, logistickú regresiu alebo rozhodovacie stromy. Výskum, ktorý realizovali Bogaert a kol. (2017), bol zameraný na cielenie online reklamy prostredníctvom sociálnej siete Facebook na užívateľov, ktorí by si mohli kúpiť produkty súvisiace s futbalom. Najskôr pomocou premenných, ktoré Facebook o užívateľoch poskytuje vyhodnotili, či užívateľ hrá alebo nehrá futbal. Následne vo svojej štúdii použili na predikciu až šesť algoritmov – model logistickej

regresie, model náhodného lesa, Adaboost, kernel factory, neurónovú sieť a model rotation forest. Z hľadiska presnosti (Accuracy) boli najlepšie modely Adaboost a model logistickej regresie. Zistené výsledky naznačovali, že premenné, ktoré boli v modeloch využité možno rozdeliť do troch kategórií: všeobecné premenné Facebooku, technologické premenné a premenné špecifické pre futbal (napr. počet priateľov, ktorí hrajú futbal, členstvo vo futbalovej skupine, počet obľúbených tímov, počet sledovaných športovcov a pod.). Vytvorené modely a výsledky z výskumu môžu využiť predajcovia na predaj futbalového vybavenia alebo futbalové tímy na zvýšenie predaja permanentiek. Výskum poukazuje na možnosť využitia dát zo sociálnej siete na oslovenie cieľovej skupiny reklamnou kampaňou na zvýšenie predaja. V práci z oblasti poisťovníctva, kde bolo potrebné vyriešiť problém s nevybalansovaným dátovým súborom, využili Lin a kol. (2017) model náhodného lesa (*random forest*) na to, aby vytvorili prediktívny model na oslovenie potenciálnych klientov (na čo najpresnejšie oslovenie cieľovej skupiny).

Pri cielení na želané publikum je možné využiť lepšie spôsoby cielenia, ako len kontextové zacielenie, cielenie na kľúčov slová, geografické cielenie alebo retargeting. Jednou z foriem cielenia je demografické cielenie, pri ktorom sa informácie získavajú väčšinou z online prieskumov (respondenti sa dobrovoľne registrujú a vypĺňajú osobné údaje, informácie o nákupnom správaní a pod., pričom sú často honorovaní za vypĺňanie týchto prieskumov). Náš výskum využíva behaviorálne cielenie a podľa Chena a Stallaerta (2014) sa online reklama, založená na behaviorálnom cielení, stáva významným odvetvím. V ich príspevku sú analyzované aj ekonomické dôsledky tohto cielenia na hlavných aktérov (zadávateľov reklamy a poskytovateľov reklamného priestoru). De Bock a Van den Poel (2010) sa vo svojej práci zamerali na využitie demografického cielenia a predikciu demografických profilov návštevníkov webovej stránky, ktoré možno použiť na účely cielenia online reklamy. Na predikciu využili model náhodného lesa. V článku je opísaná metodológia odvodzovania demografických atribútov, ako pohlavie, vek, kategórie povolania a úrovne vzdelania od anonymných návštevníkov webových stránok pomocou vzorov klikania na jednotlivé časti stránky, čo následne slúži ako vstupy pre klasifikátory (klasifikačné algoritmy) modelu náhodného lesa. Demografické profily používateľov pomáhajú marketingovým manažérom pri výbere komunikačného kanála a umožňujú personalizáciu reklamnej kampane na cieľové publikum. Extrahované informácie z webovej stránky, ktoré sa následne využijú na analýzu sú jednotlivé záujmy návštevníka na stránke, čas dňa a deň v týždni, kedy návštevník stránku navštívil a frekvencia návštev. V uvedenej práci bol model náhodného

lesa porovnávaný s algoritmami CART, C4.5, Bagging a AdaBoost. Z uvedených algoritmov vyšiel najlepšie model náhodného lesa. Získané výsledky a publiká bolo následne možné využiť pri výbere webových stránok, ktoré sa majú použiť na online reklamu. Nedostatkom tohto modelu je, že výsledky nie je možné využiť v reálnom čase.

Tieto a mnohé ďalšie príklady hovoria o dôležitosti prepojenia online marketingu s metódami strojového učenia a s oblasťou data miningu pri vytváraní rôznych druhov prediktívnych modelov.

1.1 Marketing a reklama

Dnešný svet je svetom inovácií a digitálnych technológií. Pokrok a zmeny sa prejavujú takmer vo všetkých aspektoch nášho života. Schopnosť vykonávať zmeny alebo sa prispôsobovať trendom a zmenám je dominantnou skutočnosťou v každej oblasti. Platí to najmä v marketingu, kde sa neustále zrýchľuje tempo zmien a zaujať bežného zákazníka je náročnejšie ako kedykoľvek predtým.

Potreby a priania dnešných spotrebiteľov sú rozhodujúcimi problémami, ktoré sa riešia pri vytváraní nových výrobkov a služieb a pri plánovaní stratégie na predaj týchto výrobkov a služieb so ziskom. Avšak potreba porozumieť a predvídať budúcich zákazníkov sa nevyhnutne stane ešte dôležitejšou ako v minulosti, pretože koneční používatelia produktov sa neustále líšia, či už z hľadiska charakteru alebo ich počtu (Louth, 2018).

Pre marketingové agentúry, ktoré sa snažia osloviť potenciálnych zákazníkov a udržať existujúcich je pravidlom, aby nepredpokladali, že včerajší zákazníci budú k dispozícii aj zajtra. Veľmi dôležité je mať aj dostatok informácií o trhu a ponúkaných produktoch. Pokiaľ firmy nedokážu držať krok s tým, čo sa deje na trhoch a akí zákazníci majú záujem o konkrétne druhy produktov, môže byť úsilie celej spoločnosti zamerané na nesprávnych ľudí v nesprávnom čase.

Rovnako nesprávne je pristupovať k spotrebiteľom s rovnakou stratégiou, resp. pristupovať k nim ako k „priemernému spotrebiteľovi“. Spoločnosť, ktorá nedokáže presne cieliť na potenciálnych zákazníkov, ktorí by mohli mať o jej produkt záujem, si nielen zvyšuje náklady na reklamu, ale môže potenciálnych zákazníkov aj odradiť tým, že sa bude objavovať so svojou reklamnou kampaňou nadmerne alebo v nesprávnej cieľovej skupine.

Významnou súčasťou spoznávania spotrebiteľov je využívanie marketingového výskumu. Ak vedia marketingové agentúry, zadávatelia reklamy alebo marketingové oddelenia vo firmách údaje z tohto marketingového výskumu vhodne využiť, môže im to priniesť dôležité informácie o spotrebiteľoch. V súčasnosti sa prevažná časť marketingového výskumu venuje činnostiam, ako je rozvoj trhového potenciálu (pre existujúce aj nové produkty), analýza nákupných zvykov a požiadaviek spotrebiteľov a zákazníkov, meranie efektívnosti reklamy, určovanie charakteristík trhu, analýza predaja a pod. Marketingový výskum sa môže použiť napríklad aj na zistenie najefektívnejších distribučných kanálov pre konkrétny produkt. Dalo by sa povedať, že marketingový výskum je jednoduchou a rýchlou formou, ako preniknúť a informovať sa o svete súčasných spotrebiteľov a dozvedieť sa o nich dôležité informácie, ktoré sa týkajú nákupného správania.

Marketing, tak ako bol pôvodne zamýšľaný, je dnes dôležitejší ako kedykoľvek predtým. Svet je zaplavený inovatívnymi výrobkami, službami, technológiami a riešeniami, ktoré keď prichádzajú na trh, musia byť komercionalizované, aby generovali výnos a zisk. Samotná inovácia nedokáže udržať spoločnosť, musí byť spojená s marketingom.

1.1.1 Reklama a cieľová skupina

Marketing je o prepojení správnych zákazníkov so správnym produktom. Reklama, ako jedna z foriem marketingovej komunikácie, slúži na sprístupňovanie produktu alebo služby konkrétnej značky cieľovému publiku (aj publiku vo všeobecnosti). Cieľom reklamy je doručiť hlavnú myšlienku vybranej cieľovej skupine a zároveň do istej miery ovplyvniť nákupné správanie spotrebiteľov a zákazníkov, za účelom dosiahnuť zisk, teda dosiahnuť to, aby si spotrebitelia kúpili daný produkt alebo službu. V business sektore zadávatelia reklamy používajú reklamu na oslovenie cieľovej skupiny, prostredníctvom vhodne zvolených komunikačných kanálov.

Cieľová skupina predstavuje potenciálnych zákazníkov produktu, teda skupinu ľudí, komu je produkt alebo služba adresovaná. Definovanie cieľovej skupiny je dôležitou súčasťou reklamnej kampane. Môže byť opísaná na základe behaviorálnych alebo demografických znakov, ako sú vek, pohlavie, príjem, vzdelanie, rodinný stav, nákupné zvyky, životný štýl, záujmy a pod. Nájdenie správneho publika môže byť kľúčové pre efektívnu reklamnú kampaň.

1.1.2 Personalizovaná reklama

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. Väčšina spoločností si uvedomuje, že personalizovaná reklama je žiaduca, ale len málokto si uvedomí, aká je dôležitá pri komunikácii s existujúcimi a potenciálnymi zákazníkmi. Z prieskumu vyšlo, že až 44 % spotrebiteľov je ochotných prejsť ku značke, ktorá má lepšie prispôsobenú marketingovú komunikáciu. Okrem toho, 90 % spotrebiteľov tvrdí, že reklamy a správy od značiek, ktoré nie sú pre nich osobne relevantné, sú nepríjemné a obťažujúce. 53 % z týchto respondentov by zaradilo takúto komunikáciu na prvé miesto v potenciálnom zozname obťažujúcich a nevhodných správ.

Ak sa na uvedené skutočnosti pozrieme z pohľadu generácií, 4 z 10 mileniálov (generácia Y)[2] tvrdia, že najdôležitejšou vecou, ktorú značka môže urobiť, je zabezpečiť, aby reklama zodpovedala ich záujmom. Vzhľadom na to, že v dnešnej dobe majú spoločnosti obrovské množstvo informácií a dát o spotrebiteľoch, je personalizovanie a správne zacielenie reklamy nutnosťou, ktoré môžu spoločnosť odlíšiť od konkurencie a spraviť ju oveľa úspešnejšou. Preto sme sa aj rozhodli v našej práci venovať práve cieleniu reklamy a osloveniu relevantnej cieľovej skupiny, čo zabezpečíme vytvorením prediktívnych modelov, ktoré budú čo najpresnejšie predikovať danú cieľovú skupinu, čím sa reklama stáva oveľa efektívnejšou.

1.2 Internetová populácia

Internetová (online) populácia predstavuje ľudí v danej krajine alebo regióne, ktorí používajú internet. IWS[3] definuje internetového používateľa ako každého, kto je momentálne schopný používať internet:

- osoba musí mať dostupný prístup na internet,
- osoba musí mať základné znalosti potrebné na používanie webovej technológie (Igi-global, 2022).

V súčasnosti tvorí svetovú internetovú populáciu 5 mld. ľudí, zatiaľ čo v roku 1995 to bolo menej ako 1 % (Statista, 2022).

[2] Generácia Y – podľa Wikipédie sú to ľudia narodení medzi rokmi 1981 a 1996
[3] IWS – Internet World Stats

Aj na Slovensku rastie internetová populácia každým rokom. V roku 2021 až 79 % ľudí vo veku 12 – 79 rokov pristupovalo na internet a celková internetová populácia predstavovala 3,6 mil. Slovákov, čo oproti roku 2020 predstavovalo 2 % nárast (IAB Slovakia, 2022).

1.3 Digitálny a online marketing

Digitálny marketing predstavuje dosiahnutie marketingových cieľov pomocou digitálnych technológií a médií. Medzi digitálne technológie a médiá zaraďujeme webové stránky, mobilné aplikácie, sociálne siete, vyhľadávače, reklamu, e-mail, digitálne partnerstvá s inými digitálnymi spoločnosťami a pod. Podmnožinou digitálneho marketingu je online marketing, tiež nazývaný ako internetový marketing. Už z názvu vyplýva, že online (internetový) marketing predstavuje akúkoľvek formu reklamy realizovanú prostredníctvom internetu.

1.3.1 Výhody online marketingu

Medzi najvýznamnejšie výhody online marketingu patria:

- Merateľnosť – výsledky online marketingu sú merateľné v reálnom čase, čím je možné vykonávať na získaných dátach rôzne štatistické analýzy, a teda získať presnejšie a lepšie predikcie do budúcnosti. Tým, že výsledky sú merateľné vieme veľmi ľahko zistiť, ktorá reklamná kampaň bola úspešná alebo neúspešná.
- Nákladovosť – internetová reklama je oveľa lacnejšia ako reklama v tradičných médiách, ako sú televízii alebo rádio. Z hľadiska efektívnosti to však neznamená, že by bola menej efektívna. Vhodne a efektívne naplánovaná reklamná kampaň online marketingu môže zasiahnuť relevantných internetových užívateľov za oveľa nižšie náklady ako kampaň tradičného marketingu.
- Cieľová skupina – pre online reklamnú kampaň je oveľa jednoduchšie nájsť a osloviť vybranú cieľovú skupinu na internete, pričom cielenie je možné až na úroveň jednotlivca. V prípade tradičných médií to môže byť náročnejšie, keďže napríklad pri reklame v novinách alebo v časopisoch si nevyberáme, aké publikum svojou reklamou zasiahneme. Televízia a rádio umožňujú určitý spôsob cielenia na svojich divákov, resp. poslucháčov, avšak tradičný marketing vo

všeobecnosti neposkytuje rovnaké možnosti cielenia ako online marketing. Pri online marketingu je možnosť zabezpečiť, aby si správni užívatelia prezerali náš obsah. SEO[4] umožňuje osloviť takých užívateľov, ktorí na internete vyhľadávajú obsah a témy pre nás relevantné. Okrem toho online reklama ponúka celé spektrum cielení, či už na základe demografických, behaviorálnych alebo geografických charakteristík.

- Nepretržitá dostupnosť a globálnosť – online reklama nie je obmedzená časom a je dostupná počas celého dňa kdekoľvek na svete. Takéto propagovanie produktov a služieb je výhodné pre medzinárodné firmy.
- Online nákupná cesta – v súčasnej dobe moderných technológií začína veľa ľudí svoju nákupnú cestu práve na internete. V procese rozhodovania o produkte alebo službe využívajú spotrebitelia vyhľadávače na získanie odpovedí na svoje otázky. Rovnako tak používajú vyhľadávače v prípade, kedy hľadajú informácie o konkrétnych značkách produktov alebo služieb. Toto môže byť skvelou príležitosťou pre spoločnosti na oslovenie a odchytenie potenciálnych zákazníkov, aby si pri výbere značky vybrali práve tú ich (Adil, 2022).

1.3.2 Nevýhody online marketingu

Okrem výhod sa pri online marketingu nájdu aj nevýhody:

- Bannerová slepota a ad-blocky – v dôsledku veľkého množstva online reklám na internete sa naučilo veľa zákazníkov tieto reklamy ignorovať. Preto sa môže ľahko stať, že našu reklamu prehliadnu. Okrem toho veľa ľudí používa aj tzv. ad blocks, ktoré zabraňujú zobrazovaniu reklám na webových stránkach.
- Problémy so zobrazením reklamy a obmedzený prístup na internet – pri zobrazení reklamy môže dôjsť k rôznym problémom, napríklad v dôsledku pomalého internetového pripojenia alebo v dôsledku zle nasadenej reklamy. Rovnako je problém, ak niektoré oblasti nedisponujú internetovým pripojením, čím sú pre online reklamu nezasiahnuteľné.
- Vysoká konkurencia – ak spotrebitelia hľadajú výrobky či služby na internete a konkurenčné značky využívajú rovnakú alebo podobnú marketingovú stratégiu, potom sa užívateľovi okrem našej reklamy objaví aj reklama konkurencie.

[4] SEO (*Search Engine Optimization*) – Optimalizácia pre vyhľadávače

V prípade, že je táto reklama zaujímavejšia a ponuka je výhodnejšia, zákazníka stratíme.

- Pirátstvo pri kopírovaní marketingovej stratégie – pirátstvo je veľkou nevýhodou online marketingu. Existuje veľa hackerov na sledovanie a kopírovanie marketingových stratégií.
- Negatívna spätná väzba – čo sa týka sociálnych médií, jediný príspevok, komentár alebo akékoľvek negatívne tvrdenie o značke, môže dlhodobo zničiť povesť značky (Infolearners, 2022).

1.4 Základné pojmy online marketingu

Základné pojmy, ktoré sa využívajú v online marketingu sú:

- Cielenie (*Targeting*) – nastavenie reklamnej kampane tak, aby zasiahla cieľovú skupinu, pre ktorú je relevantná, čo najpresnejšie a najefektívnejšie.
- Cieľová skupina (*Target audience*) – predstavuje skupinu ľudí (užívateľov), ktorí s najväčšou pravdepodobnosťou budú chcieť predávaný produkt alebo službu, čiže ide o skupinu ľudí, ktorí by mali vidieť cielenú reklamnú kampaň.
- Cookies – údaj z webovej stránky, ktorý je uložený vo webovom prehliadači. Používajú sa na informovanie servera, že sa používatelia vrátili na konkrétnu webovú stránku.
- CPM (*Cost per mille*) – cena za tisíc zobrazení reklamy.
- CTR (*Click Through Rate*) – percentuálna miera prekliku, ktorá sa vypočíta ako podiel klikov na reklamu a impresií, vynásobený číslom sto.
- Impresia (*Impression*) – zobrazenie online reklamy.
- Klik (*Click*) – kliknutie na online reklamu (napr. na banner).
- Konverzia (*Conversion, Action*) je definovaná ako akcia na stránke klienta (napr. vyplnenie formulára, predaj, objednanie tovaru a pod.).
- Konverzný pomer (*Conversion rate*) – percento zákazníkov alebo potenciálnych zákazníkov, ktorý vykonajú konverziu. Ide o podiel konverzií a reálnych používateľov vyjadrený v percentách.
- Obsah (*Content*) – obsah webovej stránky, o ktorý sa užívateľ alebo návštevník zaujíma.

- Reálni používatelia (*Real users*) – počet, ktorý predstavuje konkrétny počet osôb, čiže nie počet cookies, IP adries, počítačov, tabletov alebo mobilných telefónov. Jeden reálny používateľ môže mať viacero zariadení, čiže aj viac cookies alebo IP adries.

1.5 Nákup online reklamy

Súčasťou digitálneho a online marketingu je displejová reklama, pod ktorou rozumieme statické (obrázky) alebo dynamické (videá) bannery alebo ich kombinácie. Displejová reklama povzbudzuje používateľa k prekliknutiu na vstupnú stránku a na vykonanie určitej akcie (napr. nákupu).

Poznáme dva typy nákupu takejto reklamy – tradičný nákup a programatický nákup (Mindshare Slovakia, 2016).

1.5.1 Tradičný nákup

Tradičný nákup online reklamy je proces hľadania správneho poskytovateľa reklamného priestoru (publishera), vyjednávania o cene, nákupu konkrétneho reklamného priestoru a konečného zobrazenia reklamy na vybranej webovej stránke. Celý proces je manuálny a vyžaduje veľa ľudského zásahu. Reklamná agentúra, resp. zadávateľ reklamy vyplní objednávku na nákup reklamného priestoru (priestor inzercie), ktorý poskytovateľ reklamného priestoru alebo médiá vytvárajú a ponúkajú. Tento reklamný priestor, resp. webová stránka sa vyberá podľa toho, či je na tejto stránke želaná cieľová skupina afinitná (webová stránka je pre vybranú cieľovú skupinu relevantná). Po nákupe zadávatelia reklamy dúfajú, že ich cieľová skupina v období reklamnej kampane navštívi túto webovú stránku čo najčastejšie, a teda danú reklamu uvidí. Keďže zadávatelia reklamy sa zameriavajú na konkrétnu webovú stránku, výsledok tradičného nákupu má tendenciu zobrazovať reklamy všetkým návštevníkom webových stránok, čo môže byť nepríjemné pre niektorých z nich, ktorí nemajú záujem o inzerovaný produkt alebo službu a pre ktorých je reklama nerelevantná. Ako príklad uvedieme predajcu automobilov, ktorý predáva nové modely áut a umiestni svoju reklamu na webovú stránku autobazáru. Hoci je obsah stránky relevantný, reklamu uvidia aj takí užívatelia, pre ktorých nie sú ponúkané modely áut cenovo dostupné.

1.5.2 Programatický nákup

Pri programatickom nákupe nás nezaujíma konkrétna webstránka alebo reklamný priestor, ale zameriavame sa na konkrétneho užívateľa, nech sa nachádza kdekoľvek na internete. Na základe dát, ktoré môžeme o užívateľovi (resp. cookie) získať, zobrazíme reklamu len tým užívateľom, ktorí svojimi charakteristikami alebo svojím správaním zodpovedajú parametrom zadanej cieľovej skupiny. Ak máme o týchto užívateľoch informácie, že sa oni alebo im podobní užívatelia práve zaujímajú o nákup produktu, zasiahne ich naša reklama, bez ohľadu na to, kde na internete sa nachádzajú.

Médium (poskytovateľ reklamného priestoru) poskytne tento priestor buď:

- na voľný trh – *Real Time Bidding* (RTB), na ktorom vytvára ponuku, ktorá sa stretne s dopytom zadávateľov reklamy (resp. reklamných agentúr). Ide o proces podobný aukcii a v reálnom čase zadávatelia medzi sebou súťažia o konkrétne zobrazenie reklamy (impresiu) užívateľom, alebo
- si dohaduje prioritné obchody (väčšieho množstva zobrazení reklamy - impresií) s konkrétnymi zadávateľmi – *Private Deals* za vopred dohodnutých podmienok a pokiaľ nepríde od týchto zadávateľov dopyt pre konkrétne zobrazenie, tak sa toto zobrazenie reklamy posunie na voľný trh.

Výhodou programatického nákupu online reklamy je:

- menšie plytvanie zobrazeniami reklamy (impresiami), keďže sa reklama zobrazuje len relevantnému publiku, a zároveň vyššia nákladová efektivita,
- širšie možnosti cielenia (napr. podľa záujmov, kľúčových slov, zariadení, značiek mobilných telefónov a pod.),
- väčšia kontrola nad tým, kde a pri akom obsahu sa naša reklama zobrazuje,
- optimalizácia v reálnom čase (Mindshare Slovakia, 2016).

Proces programatického nákupu prebieha automatizovane a v reálnom čase. Na strane poskytovateľa reklamného priestoru (angl. *publishers*) prebieha prostredníctvom tzv. SSP (*Supply Side Platform*) a na strane zadávateľa reklamy (reklamná agentúra alebo klient) prostredníctvom tzv. DSP (*Demand Side Platform*).

DSP (*Demand Side Platform*) využívajú inzerenti (reklamné agentúry alebo reklamné siete) na nákup digitálneho priestoru. Zadávatelia majú prostredníctvom DSP prístup k ponúkanému mediálnemu priestoru. Softvér DSP sa využíva priamo v procese automatizovanej aukcie na oslovenie cieľového publika za čo najefektívnejšiu a najvýhodnejšiu cenu.

SSP (Supply Side Platform) je technológia, ktorá umožňuje ponúkať a predávať poskytovateľom mediálneho priestoru tento mediálny priestor na ich web stránkach. Rovnako ako v prípade DSP, aj tu prebieha predaj mediálneho priestoru automaticky, čo prináša poskytovateľom určité výhody, ako negociácia cien a možnosť nastaviť si podmienky predaja.

Ad Exchange predstavuje voľný trh, resp. digitálny trh reklám, na ktorom medzi sebou zadávatelia a poskytovatelia obchodujú. Celý proces prebieha automatizovane prostredníctvom aukcie v reálnom čase a predmetom obchodovania je digitálny reklamný priestor. Obe strany z tohto obchodovania profitujú, keďže poskytovatelia ponúkajú svoj reklamný priestor väčšiemu množstvu záujemcov a zadávatelia majú väčšiu škálu výberu pre nákup reklamy na rôznych webových stránkach (Mindshare Slovakia, 2018).

1.6 Typy cielenia online reklamy

V súčasnom konkurenčnom prostredí internetovej reklamy si reklamné agentúry nemôžu dovoliť míňať financie na zobrazenia reklamy užívateľom, ktorí nie sú pre klientov reklamných agentúr (zadávateľov reklamy) zaujímaví a nepatria do cieľovej skupiny, na ktorú je kampaň zameraná. Rovnako tak je na agentúry vyvíjaný neustály tlak kvôli znižovaniu nákladov na reklamu a teda na efektívne cielenie. Vyššie spomínaný programatický nákup dokáže tieto želania klientov napĺňať. Existuje niekoľko prístupov opisu cielenia online reklamy. My uvedieme prístup spoločnosti GroupM Slovakia[5]:

1. úroveň: najrozšírenejšie cielenie (*broad*) – ide o cielenie bez špecifických obmedzení, pri ktorom nie je potrebné využívať obmedzujúce podmienky. Používa sa najmä pri budovaní povedomia o značke a ide o cielenie zamerané na rozširovanie zásahu kampane.
2. úroveň: behaviorálne/profilové cielenie – cieľová skupina je identifikovaná podľa správania na internete, napr. podľa obsahu, ktorý vyhľadáva, teda či má záujem o cestovanie (navštevuje webové stránky s tematikou cestovania), alebo o šport, módu a pod.
3. úroveň: demografické cielenie – cielenie podľa demografických charakteristík, ako pohlavie, vek, rodinný stav, vzdelanie a i. Tieto informácie

[5] Tento prístup je verejne dostupný (Mindshare Slovakia, 2017) a v práci je uvedený so súhlasom spoločnosti GroupM Slovakia

sa často zisťujú prostredníctvom online prieskumov, ktoré užívatelia pri registrácii dobrovoľne vypĺňajú, pričom sú častokrát za vypĺňanie prieskumov honorovaní. Tieto informácie je v ďalších krokoch možné využiť pri tzv. Look-alike modelingu (hľadanie užívateľov s podobnými charakteristikami).

4. úroveň: geografické cielenie – údaje pochádzajú z IP adries, ktoré identifikujú impresiu (zobrazenie) v momente, keď je ponúknutá na predaj pri procese programatického nákupu, z pohľadu krajiny, okresu, regiónu a pod.
5. úroveň: kontextuálne cielenie – využíva špecifické kľúčové slová – obsah, pri ktorom je možné reklamu zobraziť.
6. úroveň: retargeting a Look-alike modeling – retargeting predstavuje identifikovanie návštevníkov konkrétnej webstránky, pričom cieľom je znovu ich v budúcnosti osloviť s novou alebo rozšírenou ponukou. Look-alike modeling hľadá užívateľov s rovnakými alebo podobnými charakteristikami, ako má návštevník stránky. Cieľom je rozšírenie osloveného relevantného publika.
7. úroveň: kombinované cielenie – rôzne kombinácie vyššie uvedených cielení. Napríklad: požiadavka pre spoločnosť, ktorá ponúka produkty osobnej starostlivosti pre ženy a zároveň chce odprezentovať v pripravovanej mediálnej online kampani otvorenie novej predajne v Bratislave, je relevantná na cieľovú skupinu: ženy žijúce v Bratislavskom kraji (alebo priamo v Bratislave a okolí), ktoré sa zaujímajú o webové stránky s produktami osobnej starostlivosti alebo beauty obsahom (Mindshare Slovakia, 2017).

Takéto a podobné požiadavky umožňuje programatický nákup online reklamy splniť aj tým najnáročnejším klientom (zadávateľom).

2 Cieľ práce

Hlavným cieľom dizertačnej práce je identifikovanie profilu internetových užívateľov, ktorí patria do cieľovej skupiny, ktorá je charakterizovaná určitými behaviorálnymi znakmi a na ktorú má byť cielená online reklama, pomocou prediktívneho modelovania. Aplikovaním metód strojového učenia vytvoríme tri prediktívne modely – model logistickej regresie (*logistic regression model*), model rozhodovacieho stromu (*decision tree model*) a model náhodného lesa (*random forest model*). Cieľom je porovnať modely a vybrať ten, ktorý predikuje cieľovú skupinu najlepšie. V práci sa zameriame na predikovanie cieľovej skupiny predstavujúcej návštevníkov internetovej stránky s online spravodajstvom, ktorí sú ochotní a majú najvyššiu tendenciu predplatiť si zamknutý obsah tejto stránky. Behaviorálne znaky sú identifikované z online správania používateľov. Vytvorené prediktívne modely budú slúžiť na optimalizáciu online kampane a na efektívne oslovenie cieľovej skupiny.

Pre naplnenie hlavného cieľa je dôležité dosiahnuť niekoľko čiastkových cieľov. Prvým čiastkovým cieľom je spracovanie aktuálnych informácií a poznatkov o digitálnom marketingu. V rámci tejto úlohy sa musíme oboznámiť so základnými pojmami z oblasti digitálneho marketingu a s výhodami a nevýhodami, ktoré tento druh marketingu so sebou prináša. Ďalej musíme pochopiť fungovanie rôznych typov online kampaní, pričom ide o typy z hľadiska nákupu online reklamy (s primárnym zameraním na programatický nákup reklamy) a typy z hľadiska cielenia online reklamy.

Druhým čiastkovým cieľom je zosumarizovať tri metódy strojového učenia (*machine learningu*), ktoré nám poslúžia na vytvorenie prediktívnych modelov. Model, ktorý bude cieľovú skupinu predikovať najlepšie, bude použitý na cielenie v reklamnej online kampani. Konkrétne sa v práci budeme venovať modelu binárnej logistickej regresie, modelu rozhodovacieho stromu a modelu náhodného lesa (*random forest*).

Tretí čiastkový cieľ spočíva v praktickej aplikácii spomínaných metód strojového učenia na údajoch z internetovej stránky. Vzhľadom na to, že prediktívne modely majú čo najlepšie vystihovať charakter dát, dôležitou súčasťou práce bude aj príprava dát, ktoré poslúžia na vytvorenie modelov. Vstupné premenné budeme detailnejšie analyzovať a okrem toho budeme prezentovať konkrétne postupy prípravy týchto premenných, aby boli vhodnými vstupnými dátami do vytváraných modelov. Po úprave vstupných údajov bude nasledovať vytvorenie troch prediktívnych modelov, čo zahŕňa aj detailný opis nastavenia parametrov a vlastností týchto modelov. Kvalitu vytvorených modelov

budeme porovnávať prostredníctvom mier využívaných na hodnotenie modelov – klasifikačných mier, ROC krivky a *F*-skóre. Na praktickú aplikáciu bude využitý analytický softvér SAS, ktorý je využívaným a veľmi dobre hodnoteným nástrojom v daných oblastiach.

Ďalším z cieľov je identifikácia faktorov, ktoré signifikantne ovplyvňujú pravdepodobnosť predplatenia si plateného obsahu na spravodajskej webstránke. Pre všetky tri modely bude vytvorený profil užívateľa, ktorý sa s najväčšou pravdepodobnosťou stane predplatiteľom plateného obsahu (zamknutých článkov na spravodajskej webstránke) a rovnako tak profil užívateľa, pri ktorom je pravdepodobnosť predplatenia si tohto obsahu veľmi nízka.

Výsledkom práce bude prediktívny model schopný predpovedať užívateľov, ktorí by sa s najvyššou pravdepodobnosťou mohli stať predplatiteľmi zamknutého obsahu spravodajskej webstránky. Keďže sa sústredíme aj na nájdenie signifikantných charakteristík existujúcich predplatiteľov, tak to nám poslúži na hľadanie nových užívateľov mimo súčasných, na ktorých môžeme následne cieliť online reklamu. Títo užívatelia budú vykazovať podobné behaviorálne znaky ako súčasní predplatitelia. Práca bude poskytovať návod a inšpiráciu ako vytvoriť a využiť prediktívny model aj pri iných webstránkach na to, aby sa z existujúcich alebo potenciálnych návštevníkov stali predplatitelia. Hoci vytvorený prediktívny model nebude univerzálny pre každú webstránku, návod a postupnosť krokov budú využiteľné aj pri iných webových stránkach a online reklamných kampaniach.

3 Metodika práce a metódy skúmania

Táto časť práce sa bude venovať základným metódam a metodickým postupom, ktoré budú v práci pri analýzach využité.

3.1 Strojové učenie – machine learning

Strojové učenie alebo tzv. *machine learning*, je v súčasnosti veľmi populárna oblasť, ktorá sa využíva v rozličných odvetviach. Dokonca by sa dalo povedať, že sa využíva všade okolo nás – v zdravotníctve, vo vzdelávaní, v zábave, pri odhaľovaní podvodov či v priemysle. Za posledné desaťročie nám strojové učenie prinieslo autonómne vozidlá, rozoznávanie hovorenej reči, efektívne vyhľadávanie na webe, personalizované reklamy, rýchlejšie diagnostikovanie chorôb v zdravotníctve, automatické generovanie titulkov a pod.

Strojové učenie je proces použitia matematických modelov a algoritmov, pomocou ktorých sa počítače učia bez priamo zadaných inštrukcií. Považuje sa za súčasť umelej inteligencie. Algoritmus vo všeobecnosti vnímame ako logickú postupnosť pokynov a krokov, ktoré musia nasledovať pri riešení problému. Algoritmy strojového učenia sú využívané k identifikácii vzorov v dátach a tieto vzory sa potom používajú k vytvoreniu dátového modelu, ktorý dokáže vytvárať prognózy. Vytvorené modely umožňujú analýzu veľkého množstva dát s rýchlymi a presnými výsledkami a predikciami, ktoré generujú rozhodnutia bez ľudského zásahu. Čím kvalitnejší je algoritmus, tým presnejšie budú rozhodnutia a predpovede pri spracovávaní údajov.

Popularita strojového učenia vzrástla vďaka nízkym nákladom na hardvér, rýchlemu spracovaniu, cloudovým technológiám, jednoduchému ukladaniu a neustále sa zvyšujúcemu objemu dát generovaných pomocou Big Data (Neto, 2020). Všetky tieto skutočnosti znamenajú, že je možné pomerne rýchlo a automatizovane vytvárať modely, ktoré dokážu analyzovať väčšie a zložitejšie údaje a poskytovať rýchlejšie a presnejšie výsledky – a to dokonca aj pri veľkých objemoch dát.

Ako sme uviedli na začiatku, dnes sú príklady strojového učenia všade okolo nás. Prvé autonómne vozidlá nás dokážu odviesť na vopred určené miesto, digitálni asistenti prehľadávajú web a dokážu reagovať na naše hlasové príkazy napríklad tým, že prehrajú hudbu. Webové stránky nám odporúčajú personalizované produkty, hudbu, filmy podľa

toho, čo sme predtým kúpili, čo sme predtým počúvali alebo pozerali. Detektory spamu zachytávajú nežiaduce správy v našich e-mailových schránkach. Zatiaľ čo my pracujeme, naše robotické vysávače vysávajú naše podlahy. Ďalší z veľkých prínosov je z oblasti medicíny – systémy analýzy obrázkov pomáhajú lekárom spozorovať nádory alebo iné ochorenia, ktoré by im mohli uniknúť.

3.1.1 Druhy strojového učenia

Vzhľadom na to, že oblasť strojového učenia je zameraná na „učenie sa“, strojové učenie sa často rozdeľuje podľa toho, ako sa algoritmus učí byť čo najpresnejší pri predikciách. Existuje niekoľko najznámejších druhov strojového učenia (Brownlee, 2019):

- Učenie sa s učiteľom (*Supervised learning*),
- Učenie sa bez učiteľa (*Unsupervised learning*),
- Učenie sa formou odmeňovania (*Reinforcement learning*),
- Učenie sa s čiastočným dohľadom (*Semi-supervised learning*),
- Učenie sa pod vlastným dohľadom (*Self-supervised learning*),
- Viacstupňové učenie sa (*Multi-Instance learning*),
- Induktívne učenie sa (*Inductive learning*),
- Deduktívne učenie sa (*Deductive learning*),
- Transdukčné učenie sa (*Transductive learning*).

Aby sme pochopili výhody a nevýhody jednotlivých druhov strojového učenia, musíme sa najskôr pozrieť na to, aké údaje do nich vstupujú. V strojovom učení existujú dva druhy údajov – označené údaje (labeled data) a neoznačené údaje (unlabeled data).

Označené údaje majú vstupné aj výstupné parametre v úplne strojovo čitateľnej podobe, ale na začiatku je potreba veľa ľudskej práce na ich označenie. Neoznačené údaje majú buď iba jeden alebo žiadny z parametrov v strojovo čitateľnej podobe. Z toho vyplýva, že na začiatku nie je potrebná žiadna ľudská práca na ich označenie, ale práca s neoznačenými dátami si vyžaduje zložitejšie riešenia.

Uviedli sme niekoľko typov strojového učenia, v súčasnosti sa však používajú najmä tri hlavné metódy: učenie sa s učiteľom, učenie sa bez učiteľa a učenie sa formou odmeňovania.

Učenie sa s učiteľom

Učenie sa s učiteľom (*supervised learning*) je jedným z najzákladnejších druhov strojového učenia. Pre množinu vstupných údajov je jasne definovaný správny výstup. V tomto type je algoritmus strojového učenia trénovaný na označených dátach. Proces učenia sa s učiteľom zahŕňa vstupné premenné, ktoré označíme X_i a výstupnú premennú, ktorú označíme Y. Na trénovacích dátach sa algoritmus naučí funkciu mapovania vstupných a výstupných (cieľových) premenných, teda aby vedel pre daný vstup vyprodukovať správny výstup. Z pohľadu štatistiky je výstup Y závislá premenná a vstup X je nezávislá premenná:

$$Y = f(X) \tag{1}$$

Naším konečným cieľom je pokúsiť sa aproximovať funkciu mapovania (f), aby sme mohli predikovať výstupnú premennú Y v prípade, že máme nové vstupné údaje X (Yagcioglu, 2020). Algoritmu strojového učenia sa na začiatku poskytne trénovací dátový súbor. Tento trénovací súbor je súčasťou väčšieho (výsledného) dátového súboru a slúži na to, aby algoritmus získal základnú predstavu o probléme a riešení. Trénovací súbor je svojimi vlastnosťami veľmi podobný výslednému dátovému súboru a poskytuje algoritmu označené parametre potrebné pre vyriešenie daného problému. Algoritmus potom nájde vzťahy medzi danými parametrami – vytvorí vzťah príčin a následkov medzi premennými v trénovacom dátovom súbore. Na konci trénovania má algoritmus predstavu o tom, aký je vzťah medzi vstupnými a výstupnými údajmi (Anirudh, 2019). Model sa upravuje (trénuje) tak dlho na trénovacej množine údajov, až kým výsledky nie sú čo najpodobnejšie skutočnosti.

Získané riešenie je potom aplikované na výsledný dátový súbor. S týmto súborom algoritmus pracuje podobne ako s trénovacím súborom, čo znamená, že daný algoritmus strojového učenia sa bude aj naďalej učiť a zlepšovať, objavovať nové vzorce a vzťahy počas toho, ako je používaný na nových dátach. Cieľom učenia sa s učiteľom je vytvoriť taký algoritmus, ktorý vie spracovať aj dátové súbory, ktoré nikdy predtým nevidel.

Existujú dva základné typy problémov učenia sa s učiteľom: klasifikácia, ktorá predikuje označenie triedy, do ktorej daný údaj patrí a regresia, ktorá spočíva v predikovaní číselnej hodnoty. Klasifikácia aj regresia môžu mať jeden alebo viac vstupných premenných a tieto môžu byť numerické aj kategoriálne (Brownlee, 2019).

Klasifikácia predstavuje proces hľadania alebo objavovania modelu (funkcie), ktorá pomáha pri rozdeľovaní prípadov do viacerých preddefinovaných tried výstupnej

premennej. Výsledkom klasifikácie je funkcia, ktorá klasifikuje jednotlivé prípady do dvoch rozličných tried. Do jednej z tried umožňuje zaradiť aj nové prípady, v ktorých hodnoty výstupnej premennej nepoznáme (Terek, Horníková, Labudová, 2010). Prípady v každej triede sú navzájom podobné, ale od prípadov v ostatných triedach sa odlišujú. Príkladom klasifikácie je príklad z oblasti počítačového videa pri klasifikácii obrázkov. Cieľom je predikovať, do akej triedy obrázok patrí. Pri tomto probléme nás zaujíma, či je daný obrázok napr. automobil, lietadlo alebo nejaké zviera. Vstupom sú teda obrázky a výstupom je názov danej kategórie, do ktorej obrázok patrí.

Medzi najznámejšie algoritmy učenia sa s učiteľom patria lineárna regresia, logistická regresia, zovšeobecnený lineárny model (*Generalized Linear Model*), Bayesovský klasifikátor, rozhodovacie stromy, náhodný les (*Random Forest*), neurónové siete, metóda podporných vektorov – SVM (*Support Vector Machines*) alebo boosting s najznámejším modelom XGBoost (*eXtreme Gradient Boosting*).

Učenie sa bez učiteľa

Učenie sa bez učiteľa (*unsupervised learning*) je druh strojového učenia opisujúci skupinu problémov, ktoré zahŕňajú použitie modelu na opísanie alebo extrakciu vzťahov v dátach. Pri učení sa bez učiteľa máme iba vstupné dáta (konkrétne ide o neznačené dáta) a žiadne zodpovedajúce výstupné premenné. Cieľom je dozvedieť sa o dátach viac informácií.

Predstavte si, že ste v cudzej krajine a napríklad navštevujete trh s potravinami. Uvidíte stánok, ktorý predáva ovocie, ktoré ale nemôžete identifikovať a názov tohto ovocia nepoznáte. Môžete sa však spoľahnúť na svoje vlastné vypozorované zistenia (údaje), ktoré môžete použiť na posúdenie a vyvodenie záverov. V takomto prípade môžete ovocie ľahko oddeliť od ponúkanej zeleniny alebo iného jedla tým, že identifikujete jeho rôzne vlastnosti, ako je farba, tvar alebo veľkosť (Yagcioglu, 2020). Takto funguje učenie sa bez učiteľa.

Medzi najznámejšie problémy učenia sa bez učiteľa patria zhlukovanie, asociácia a zníženie dimenzií. O zhlukovaní (*clustering*) hovoríme vtedy, keď chceme objaviť vnútorné (inherentné) zoskupenia v dátach, napríklad zoskupiť zákazníkov podľa ich nákupného správania. Algoritmy zhlukovania spracujú naše údaje a roztriedia ich do prirodzených skupín (zhlukov), ak v údajoch existujú. Môžeme tiež upraviť, koľko zhlukov by mal náš algoritmus identifikovať. Cieľom je nájsť zhluky a interpretovať pomocou nich vstupné údaje. Zhlukovanie sa bežne používa na určovanie segmentov

zákazníkov v marketingových dátach. Schopnosť určiť rôzne segmenty zákazníkov pomáha firmám pristupovať k týmto zákazníckym segmentom jedinečným spôsobom.

Pri asociačných problémoch chceme nájsť pravidlá, ktoré opisujú veľkú časť našich údajov a vzťahy medzi nimi, napríklad ľudia, ktorí kupujú produkt A majú tendenciu kupovať produkt B.

Zníženie dimenzionality je bežne používanou technikou učenia sa bez učiteľa, ktorej cieľom je znížiť počet uvažovaných náhodných premenných. Jedným z najbežnejších spôsobov jej použitia je zníženie zložitosti problému premietnutím priestoru funkcií do priestoru s nižším počtom dimenzií, aby sa v systéme strojového učenia brali do úvahy menej korelované premenné. Príkladom je metóda analýzy hlavných komponentov (PCA – *Principal Component Analysis*).

Najznámejšie algoritmy učenia sa bez učiteľa, okrem spomínanej analýzy hlavných komponentov, sú K-means pre zhlukovanie alebo Apriori algoritmus pre tvorbu asociačných pravidiel (Brownlee, 2016).

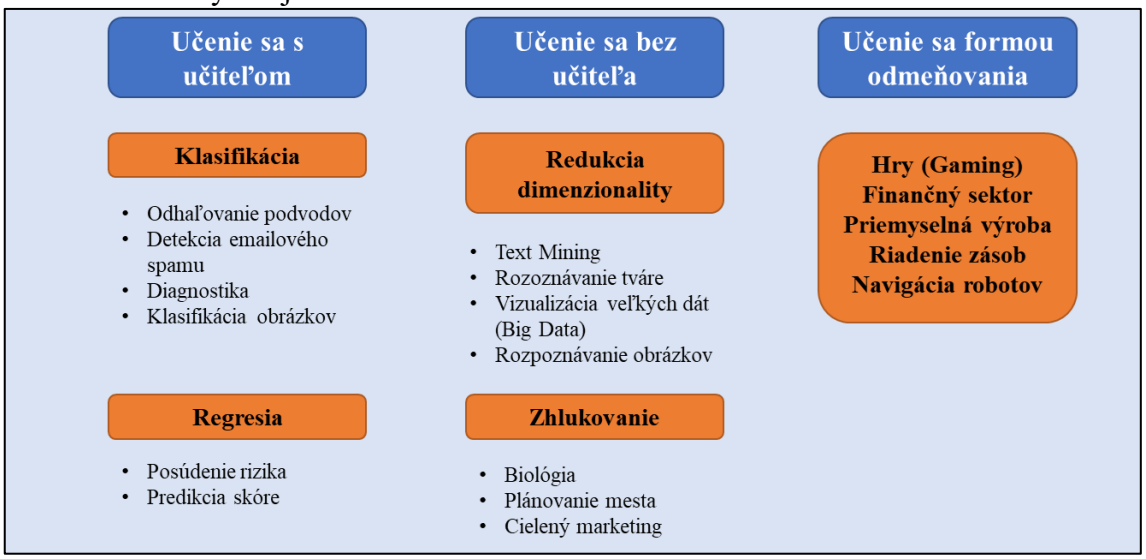
Učenie sa formou odmeňovania

Učenie sa formou odmeňovania (*reinforcement learning*) opisuje skupinu problémov, kde v danom prostredí pracuje tzv. agent a ten sa musí naučiť pracovať pomocou spätnej väzby, resp. sa učí prostredníctvom interakcie s prostredím. V tomto prípade nevystupuje trénovacia množina označených alebo neoznačených dát, avšak agentovi musíme určiť pravidlá, ako sa môže v danom prostredí správať. Okrem pravidiel musíme určiť aj odmeňovaciu funkciu, pomocou ktorej vie vyhodnotiť, či práve vykonané rozhodnutie preňho bolo alebo nebolo prospešné. Celé učenie prebieha metódou „pokus-omyl“, kde sa agent učí, ako sa má v jednotlivých situáciách ideálne správať. Medzi algoritmy učenia sa formou odmeňovania patria napr. Q-learning alebo Deep Q-learning.

Učenie sa prostredníctvom odmeňovania ilustrujme na hre šach. Vytvoríme agenta, ktorému najskôr definujeme pravidlo pre výhru a povolené ťahy. Ak vyhrá alebo vyradí súperovu figúrku, dostane odmenu, ak však prehrá alebo mu súper vyhodí figúrku, bude potrestaný. Necháme ho trénovať, aby si zahral niekoľko miliónov partií sám proti sebe. Ako výsledok dostávame umelú inteligenciu, ktorá je veľmi dobrá v hre šach (Muráň, 2019).

Pre zhrnutie sa môžeme pozrieť na príklady využitia týchto druhov strojového učenia v praxi (Obr. č. 1).

Obr. č. 1 – Druhy strojového učenia



Zdroj: Data-flair, 2020

3.2 Text Mining

S digitálnou transformáciou celého sveta prišlo k explózii textových informácií zo širokého spektra zdrojov. S týmto príchodom došlo rovnako tak k vzostupu dolovania údajov (data mining). Najprirodzenejšou formou uchovávania informácií je text. Text mining je typ analýzy údajov, ktorej cieľom je získať cenné poznatky z textových informácií. Čo sa týka druhov strojového učenia, patrí do kategórie učenia sa bez učiteľa. Ide o proces transformácie neštruktúrovaného textu do štruktúrovaného formátu s cieľom identifikovať zmysluplné vzory a získať z textu nové poznatky.

Proces Text Miningu je zložený z niekoľkých činností. Prvou z nich je zber textových údajov, z ktorých sa vytvorí tzv. korpus – zbierka dokumentov, ktoré predstavujú vstupné údaje pre model. Zozbierané údaje sa následne môžu ukladať v databázach. Text je jedným z najbežnejších typov údajov v databázach. V závislosti od databázy môžu byť tieto údaje štruktúrované, neštruktúrované alebo pološtruktúrované:

- Štruktúrované údaje sú štandardizované do tabuľkového formátu s množstvom riadkov a stĺpcov, čo uľahčuje ukladanie a spracovanie pre analýzu. Medzi tieto údaje patria napr. mená, adresy alebo telefónne čísla.
- Neštruktúrované údaje nemajú vopred definovaný formát. Môžu obsahovať text z rôznych zdrojov, ako sociálne médiá, recenzie produktov alebo multimediálne formáty – video alebo zvukové súbory.

- Položštruktúrované údaje sú zmesou štruktúrovaných a neštruktúrovaných dátových formátov. Hoci majú určitú organizáciu, nemajú dostatočnú štruktúru na splnenie požiadavky relačnej databázy. Medzi príklady patria súbory XML, JSON a HTML (IBM Cloud Education, 2020).

3.2.1 Spracovanie textu

Skôr ako budeme môcť použiť rôzne techniky Text Miningu, musíme začať s predbežným spracovaním textu, čo zahŕňa rozbor textu, čistenie a transformáciu textových údajov do použiteľného formátu. Tento postup zvyčajne zahŕňa použitie techník, ako je identifikácia jazyka, tokenizácia – rozčlenenie textu na jednotlivé výrazy (tokeny), normalizácia týchto výrazov (napr. prostredníctvom lematizácie – angl. *lemmatization* alebo hľadania koreňa slova – angl. *stemming*), označovanie časti reči (*part-of-speech tagging*) – vyznačovanie slovných druhov. Techniky opíšeme v ďalších podkapitolách.

Existujú dva rozšírené prístupy k analýze textu. Prvá je metóda *bag-of-words,* kde základným predpokladom je počítanie slov v texte a pochopenie toho, ako tieto slová spolu syntakticky (štrukturálne) súvisia s ohľadom na gramatické pravidlá a pod. Táto metóda je postačujúca na sumarizáciu a klasifikáciu textových dokumentov, avšak poradie jednotlivých analyzovaných slov je pri tejto metóde irelevantné a neberie sa do úvahy sémantický význam jednotlivých slov.

Druhý prístup je lingvistický a predpokladá, že na skutočné pochopenie a klasifikáciu textu musíme prejsť od syntaxe (štruktúry) k sémantike (významu slov). Aby bolo možné použiť ktorýkoľvek prístup, textový dokument najskôr rozoberieme pomocou syntaktickej analýzy. Syntaktická analýza (angl. *parsing*) je jedným z najdôležitejších krokov Text Miningu. Je prvým krokom v prevedení neštruktúrovaného textu do štruktúrovaného (tabuľkového) formátu na uľahčenie analýzy (Chakraborty, Pagolu, Garla, 2013).

3.2.2 Tokenizácia – rozčlenenie textu na výrazy

Súčasťou syntaktickej analýzy je tzv. tokenizácia. Je to proces, kedy sa text rozdelí na rôzne zhluky znakov, ktoré sú oddelené medzerami alebo inými znakmi (bodka,

bodkočiarka a pod.). Vzniknuté časti – zhluky sa nazývajú tokeny a môžu zahŕňať nielen slová, ale aj frázy alebo celé vety (Mohler, 2020).

Ďalším krokom je normalizácia tokenov, keďže sa v texte nachádzajú v rôznych formách. Zahŕňa ich zmenu na základný tvar pomocou identifikácie koreňa slova. Pri hľadaní koreňa slova sa slovo upravuje na tvar, ktorý nie je skloňovaný ani časovaný, tzn. že podstatné mená sa upravia na tvar nominatívu jednotného čísla, slovesá na neurčitok, odstránia sa prípony a predpony slov. Využívajú sa na to dve metódy – lematizácia (*lemmatization*) a tzv. hľadanie koreňa slova – *stemming*. Obe metódy majú rovnaký cieľ – úpravu slovného druhu na jeho slovný základ (SAS Support, 2012). Pri stemmingu sa používa heuristický spôsob úpravy, a to odseknutie prípony zo skloňovaného slova. Oproti lematizácii nezohľadňuje slovný druh, čo môže viesť k následnej nesprávnej klasifikácii slova. Lematizácia na úpravu využíva štandardnú slovnú zásobu a morfologickú analýzu slov na redukciu skloňovaného slova do základnej alebo slovníkovej formy (nazýva sa aj lemma).

Základnou myšlienkou je, že pomocou zníženia početnosti odlišných tvarov slov v korpuse sa zjednodušuje model bez straty významných informácií. Jedným z príkladov je využitie koreňového slova „beh“ na vyjadrenie výrazov ako „bežal“, „beží“, „bežec“ a pod. Ďalším príkladom je použitie jednotného čísla podstatného mena ako koreňového slova pre množné číslo, napr. slovo „mačka“ pre množné číslo „mačky“ (Balodi, 2020). Okrem toho možno pravidlo rozšíriť aj na sémanticky ekvivalentné slová alebo synonymá, napr. lematizácia založená na synonymách môže použiť koreňové slovo „auto“ pre sémanticky ekvivalentné slovo, ako je „automobil“. Ďalej môžeme kombináciou podstatných a prídavných mien vytvoriť združené pomenovania, ako napr. „mobilný telefón“, „vysokoškolský diplom“, „oxid uhličitý“ a pod.

3.2.3 Part-of-Speech (POS) – identifikácia slovných druhov

Ďalšou dôležitou súčasťou Text Miningu je identifikácia jednotlivých tokenov z hľadiska slovných druhov na základe významu slova a kontextu. Táto úloha je veľmi náročná, pretože jedno slovo môže reprezentovať niekoľko slovných druhov v závislosti od kontextu. Napríklad slovo „mať“, môže byť aj podstatné meno (matka) aj sloveso (vlastniť). Rovnako treba dávať pozor na kontextový význam slov – napr. slovo jazyk má viacero významov, a to ľudský jazyk v ústach, jazyk ako reč – napr. slovenský, nemecký, jazyk v topánke alebo snehové jazyky. Na vyriešenie nejednoznačností a správnu

identifikáciu slovného druhu je potrebné porozumenie susedným výrazom vo vete alebo v odseku. Algoritmy POS (*Part-of-Speech*) sú vytvorené tak, aby na základe definovaných pravidiel porozumeli vzťahom medzi výrazmi. Určujú, či je slovo všeobecným alebo vlastným podstatným meno, prídavným menom, slovesom, príslovkou atď. Cieľom je vytvoriť skupiny syntakticky podobných výrazov (SAS Support, 2012).

3.2.4 Štart a stop zoznamy

S počtom výrazov využívaných v analýze sa zvyšuje aj náročnosť úlohy Text Miningu. V každom texte sa nachádzajú výrazy, ktoré sú nevýznamné pre analýzu. Väčšinou ide o slová ako „a“, „v“, „že“ a pod., pričom tieto výrazy môžu tvoriť až 50 % zo všetkých výrazov. Jednou z možností je vynechať konkrétne slovné druhy, ako sú predložky, spojky a častice. Ďalšou z metód filtrovania výrazov je využívanie tzv. štart a stop zoznamov. Tieto zoznamy pomáhajú lepšie kontrolovať výrazy, ktoré sú neskôr využité v analýze. Štandardné stop zoznamy pozostávajú najčastejšie z častíc, spojok a predložiek. Vlastné stop zoznamy umožňujú vynechanie konkrétnych výrazov, napríklad slovo smartfón v článku o smartfónoch. Opačný prípad je štart zoznam, ktorý obsahuje výrazy, ktoré určite chceme v analýze ponechať. Tie bývajú často využívané v prípadoch, ak sa v dokumentoch nachádza technický žargón.

3.2.5 Filtrovanie textových výrazov

Pri filtrovaní textových výrazov je potrebné určiť nastavenia váh výrazov a vážených frekvencií. Cieľom analýzy Text Miningu je odlíšenie dokumentov v korpuse. Kľúčovou úlohou je identifikácia významných pojmov, ktoré odlišujú jednotlivé dokumenty. Výskumy ukázali, že využívanie jednoduchých početností slov je nedostatočné pri odlišovaní jednotlivých dokumentov. Poznáme niekoľko techník pre výpočet početností výrazov, početností dokumentov, v ktorých sa výraz vyskytuje alebo veľkosti korpusu, od ktorej sa odvodzujú váhy pre každý výraz. Výrazy, ktoré sa v korpuse vyskytujú často, majú priradené vyššie váhy a sú považované za dôležité, pretože lepšie opisujú jednotlivé dokumenty. Výrazy, ktoré sa vyskytujú menej často, majú tak isto priradené vyššie váhy, pretože lepšie odlišujú dokumenty v korpuse.

Frekvenčná matica je zostavená pomocou dvoch typov váh. Prvá je váha početnosti – lokálna váha, ktorá je výsledkom transformácie početnosti výskytov výrazu

v dokumente za použitia váženej funkcie. Druhá je váha výrazu – globálna váha, ktorá je priradená každému výrazu podľa celkovej početnosti a početnosti dokumentov. Hodnota bunky v matici je váženou hodnotou početnosti získaná vynásobením lokálnej a globálnej váhy.

Jedna z najznámejších váhových schém je tzv. váha tf-idf (*term frequency, inverse document frequency,* Stecanella, 2019). Podľa nej vyrátame početnosť termínu (*tf*), teda lokálnu mieru nasledovne:

$$tf = \frac{tdf}{\max tdf} \tag{2}$$

kde tdf je početnosť špecifickosti dokumentu pre daný výraz a $\max tdf$ je početnosť najčastejšie sa vyskytujúceho výrazu v dokumente.

Globálnu mieru idf vypočítame nasledovne:

$$idf = \ln\left(\frac{N}{df_t}\right) \tag{3}$$

kde N je veľkosť zbierky dokumentov a df_t je počet dokumentov, v ktorých sa vyskytuje výraz t.

Poznáme rozličné nastavenia vážených lokálnych početností. SAS Text Miner využíva tri typy nastavení: Binary, Log a None.

$$\boldsymbol{Binary}:\ g(f_{ij}) = 1 \tag{4}$$

- Táto rovnosť platí, ak sa výraz vyskytuje v dokumente. Ak sa v dokumente nevyskytuje, potom $g(f_{ij}) = 0$. Toto nastavenie však neodlišuje dokumenty, v ktorých sa výraz vyskytuje veľmi často a dokumenty, v ktorých sa vyskytuje zriedka.
- Log nastavenie sa využíva na kontrolu účinku vysoko početných výrazov v dokumente.

$$\boldsymbol{Log}:\ g(f_{ij}) = \ln(f_{ij} + 1) \tag{5}$$

- Pri nastavení None sa používa početnosť bez vykonania akejkoľvek transformácie. Neznamená to však, že by nebola použitá žiadna metóda vážených početností.

$$\boldsymbol{None}:\ g(f_{ij}) = f_{ij} \tag{6}$$

Lokálne početnosti pomáhajú iba pri pochopení skladby textového dokumentu, ale nedokážu identifikovať pojmy, ktoré by rozlišovali dokumenty. Túto úlohu plnia globálne početnosti a každému výrazu je priradená váha pomocou niektorej z troch rôznych metód dostupných v SAS Text Mineri – entropia, vzájomná informácia a inverzná frekvencia dokumentu (*entropy, mutual information, IDF*). Poslednou možnosťou je nevyužiť žiadne váhy termínov (w_i = 1) (Chakraborty, Pagolu, Garla, 2013).

3.2.6 Zhlukovanie (Clustering)

Zhlukovanie alebo zhluková analýza je súbor techník, ktoré sa snažia nájsť prirodzené zoskupenia v údajoch. Zásadným rozdielom medzi zhlukovou analýzou a typickým klasifikačným modelom je absencia akejkoľvek cieľovej (vysvetľovanej) premennej. Cieľom je zoskupiť objekty do skupín tak, aby objekt v každom zhluku bol podobný ostatným v danom zhluku, a zároveň objekty v rôznych zhlukoch boli odlišné. V kontexte textových údajov sú objektami dokumenty, ktoré musia byť zaradené do jednotlivých zhlukov, takže v rámci jedného zhluku sú dokumenty podobné, ale medzi ostatnými zhlukmi sú dokumenty odlišné. Miery podobnosti sú v štatistike definované pomocou troch metód: vzdialenosť, asociácia a korelácia. Metódy založené na vzdialenosti a korelácii zvyčajne vyžadujú metrické údaje[6], zatiaľ čo asociačné miery fungujú na nemetrických údajoch.

Pri metódach založených na vzdialenosti teda platí, že podobnosť objektov je reprezentovaná malou vzdialenosťou. Ak sú objekty odlišné, vzdialenosť medzi nimi je veľká. Najčastejšie používané metriky vzdialenosti sú Euklidovská vzdialenosť a Mahalanobisova vzdialenosť, ktorá je definovaná nasledovne:

$$d(\mathbf{X_i}, \mathbf{X_j}) = \sqrt{(\mathbf{X_i} - \mathbf{X_j})^{\mathrm{T}} \sum^{-1} (\mathbf{X_i} - \mathbf{X_j})} \tag{7}$$

kde $\mathbf{X_i}$ a $\mathbf{X_j}$ sú vektory hodnôt premenných pre prípady i a j a $\mathbf{\Sigma}$ je kovariančná matica. Mahalanobisova vzdialenosť pri výpočte, na rozdiel od Euklidovskej, berie do úvahy aj veľkosť kovariancie medzi premennými. Pri využití SAS Text Miner sa na

[6] Metrické údaje – všetky údaje, ktoré možno merať na stupnici (škále) a môžu nadobudnúť akúkoľvek hodnotu na tejto stupnici.

meranie vzdialenosti používa Mahalanobisova vzdialenosť pri EM algoritme (*expectation-maximization algorithm*) a Euklidovská vzdialenosť pri hierarchickom zhlukovaní (Chakraborty, Pagolu, Garla, 2013).

3.2.7 Algoritmy zhlukovania

Vo všeobecnosti možno rozdeliť algoritmy zhlukovania do štyroch skupín: hierarchické algoritmy, nehierarchické algoritmy, pravdepodobnostné algoritmy a neurónové siete.

Algoritmus hierarchického zhlukovania iteratívne zoskupuje objekty do kaskádovitých skupín zhlukov. Zhluky v rámci daného kroku sú vnorené do zhluku z predchádzajúceho (alebo nasledujúceho) kroku. SAS Text Miner využíva v rámci hierarchického zhlukovania Wardovu metódu minimálneho rozptylu (Ward, 1963) na výpočet vzdialenosti medzi dvoma zhlukmi.

Nehierarchické algoritmy, ako napríklad *k*-means sú neinkrementálne[7] a pozorovania sa do zhlukov priraďujú simultánne na základe vzdialenosti každého pozorovania od stredu zhluku. Názov *k*-means znamená, že algoritmus rozdelí množinu údajov do *k* zhlukov a priradí pozorovania k centru zhluku, ktorý je najbližšie (Dalhousie University, 2012). V algoritme prebieha iteratívny proces, pri ktorom sa priraďujú pozorovania do najbližších stredov a zároveň sa tieto stredy aktualizujú, kým nie je dosiahnuté konvergenčné kritérium (Gao, Buffalo, 2012).

Pravdepodobnostné zhlukovanie identifikuje oblasti s vysokou hustotou v rámci dátového priestoru. Algoritmus, ktorý sa využíva pri tomto type zhlukovania, využíva funkciu hustoty premennej v rámci každého *k* zhluku. Najznámejší z algoritmov je EM algoritmus – algoritmus maximalizácie očakávania (*expectation-maximization algorithm*) a pozostáva z dvoch krokov. V prvom kroku sa každému pozorovaniu priradí váha alebo pravdepodobnosť výskytu v danom zhluku. V druhom kroku sa počiatočné hodnoty rozdelenia zmenia na vážený priemer priradených pozorovaní, kde sa využijú váhy identifikované v prvom kroku. Proces sa opakuje, až dokým nevykáže štatistika vierohodnosti signifikantnú zmenu v rozdelení hodnôt.

[7] Inkrementálne algoritmy spracovávajú prípady z trénovacej množiny postupne, jeden za druhým. Neinkrementálne algoritmy spracovávajú celú trénovaciu množinu odzadu. Využívajú sa na riešenie úloh, pri ktorých sú trénovacie príklady dostupné ešte pred začatím účenia.

Hoci sa neurónové siete z veľkej časti využívajú ako metóda učenia sa s učiteľom, je možné ich využiť aj pri zhlukovaní. Ide o samo-riadiacu mapu (SOM, *self-organizing map*) nazývanú aj Kohonenova mapa. Často sa využíva v rámci metód učenia sa bez učiteľa alebo pri zhlukovaní. Obsahuje zvyčajne dve vrstvy. Vstupná vrstva pozostáva z p-rozmerných pozorovaní. Výstupná vrstva pozostáva z k uzlov pre k zhlukov, z ktorých každý je spojený s p-rozmernou váhou. Každému výstupnému uzlu sa na začiatku priradí náhodná váha. Tieto náhodné váhy sa upravujú počas procesu, kedy sa Kohonenova mapa učí identifikovať vzory zo vstupných údajov. Každé vstupné pozorovanie sa dočasne priradí jednému z uzlov na základe najkratšej vzdialenosti medzi každým pozorovaním a každým uzlom. Následne sa váhy v uzloch prepočítavajú na základe toho, ktoré vstupné pozorovania sú priradené každému uzlu. Tento iteratívny proces sa opakuje dovtedy, kým váhy nezodpovedajú stredom zhlukov tak, že zhluky, ktoré sú si navzájom podobné, sú umiestnené v tesnej blízkosti v Kohenenovej mape (Chakraborty, Pagolu, Garla, 2013).

Medzi autorov, ktorí sa venovali URL[8] zhlukovaniu (URL clustering) a analýze text miningu, patria napríklad Nagwani (2010), Poomagal and Hamsapriya (2011), Zhang M. et al. (2015), Li B. et al (2019), Hernández (2012) a pod.

3.3 Logistická regresia

Logistická regresia, ako jedna z metód strojového učenia, vychádza zo zovšeobecneného lineárneho modelu (GLM – *Generalized Linear Model*). Vysvetľovaná (závislá) premenná v modeli je kategoriálna (dichotomická, viackategoriálna nominálna alebo ordinálna) a vysvetľujúce (nezávislé) premenné môžu byť kategoriálne alebo spojité. Zovšeobecnený lineárny model GLM je tvorený systematickou (deterministickou) zložkou, ktorá špecifikuje funkciu nezávislých premenných, náhodnou zložkou $Y = (Y_1, Y_2, \dots, Y_n)$, ktorá identifikuje závisle premennú a jej rozdelenie a väzbovou funkciou opisujúcou vzťah medzi systematickou zložkou a náhodnou zložkou.

Pre prípad zovšeobecneného lineárneho modelu, v ktorom má závislá premenná alternatívne rozdelenie pravdepodobnosti a väzbová funkcia je funkcia logit, obsahuje

[8] URL (*Uniform Resource Locator*) – hovorovo nazývané aj webová adresa je adresa daného jedinečného zdroja na webe.

náhodná zložka postupnosť náhodných premenných Y_i s alternatívnym rozdelením pravdepodobnosti s parametrom $\pi(\mathbf{x}_i) = \pi_i$, ktorého hodnota závisí od hodnôt i-teho riadku matice **X**:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \tag{8}$$

kde x_{ij} $(i = 1, 2, \dots, n; j = 1, 2, \dots, k)$ sú hodnoty nezávisle premenných $X_1, X_2, \dots, X_k$ zistené na prípadoch $O_1, O_2, \dots, O_n$

Pre strednú hodnotu $E(Y_i)$ a rozptyl $D(Y_i)$ premennej Y_i s alternatívnym rozdelením pravdepodobnosti platí (Terek, Horníková, Labudová, 2010):

$$E(Y_i) = E(Y|\mathbf{x}_i) = \pi_i \,.\, 1 + (1 - \pi_i) \,.\, 0 = \pi_i \tag{9}$$

$$D(Y_i) = D(Y|\mathbf{x}_i) = \pi_i \,.\, (1 - \pi_i) \tag{10}$$

Systematickou (deterministickou) zložkou GLM modelu je vektor $\eta = (\eta_1, \eta_2, \dots, \eta_n)$, ktorý nazývame aj lineárny prediktor. Vyjadrením lineárneho prediktora v maticovom tvare dostaneme:

$$\boldsymbol{\eta}^{\mathrm{T}} = \mathbf{X}\boldsymbol{\beta}^{\mathrm{T}} \tag{8}$$

kde **X** je matica, ktorá je vyjadrená vzťahom (8) a **β** je $(k + 1)$-prvkový vektor neznámych parametrov modelu.

Ak dosadíme do uvedeného vzťahu jednotlivé matice a vektory dostaneme:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \,.\, \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{12}$$

Väzbová funkcia g, ktorá je jednou zo zložiek GLM modelu, je funkciou podmienenej strednej hodnoty závislej premennej $E(Y|\mathbf{x}_i)$ a vyhovuje nasledujúcej požiadavke:

$$g\big(E(Y|\mathbf{x}_i)\big) = g(\mu_i) = g(\pi_i) = \eta_i, i = 1, 2, \dots, n \tag{13}$$

kde $\mathbf{x}_i$ je vektor hodnôt vysvetľujúcich premenných.

Väzbová funkcia, ktorá je najčastejšie používanou funkciou pre prípad závislej premennej s alternatívnym rozdelením pravdepodobnosti, je funkcia **logit**, ktorá transformuje parameter p_i na logaritmus šance (Vojtková a Stankovičová, 2020).

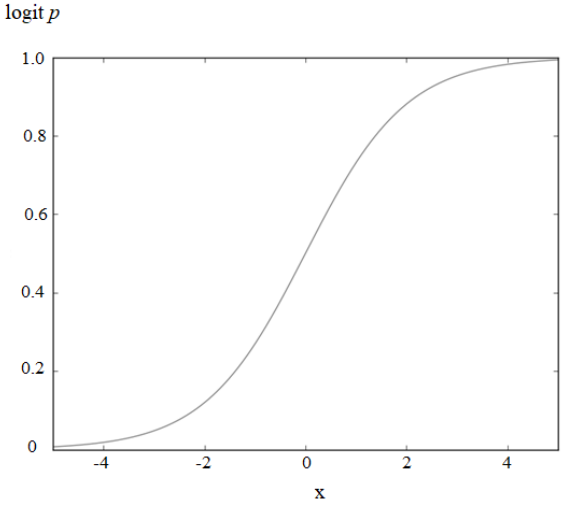
$$logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \qquad (14)$$

Rovnica modelu logistickej regresie má nasledujúci tvar:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} = \sum_{j=0}^{k} \beta_j x_{ij} \qquad (15)$$

Funkcia logit transformuje hodnoty parametra π_i, ktoré sú v intervale (0; 1) na hodnoty logaritmu šance $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$, ktoré sú z intervalu $(-\infty;\ +\infty)$.

Graf č. 1 – Graf funkcie logit



Zdroj: Vlastné spracovanie

Logaritmická funkcia je inverznou funkciou k exponenciálnej funkcii. Zo vzťahu (15) môžeme vyjadriť šancu:

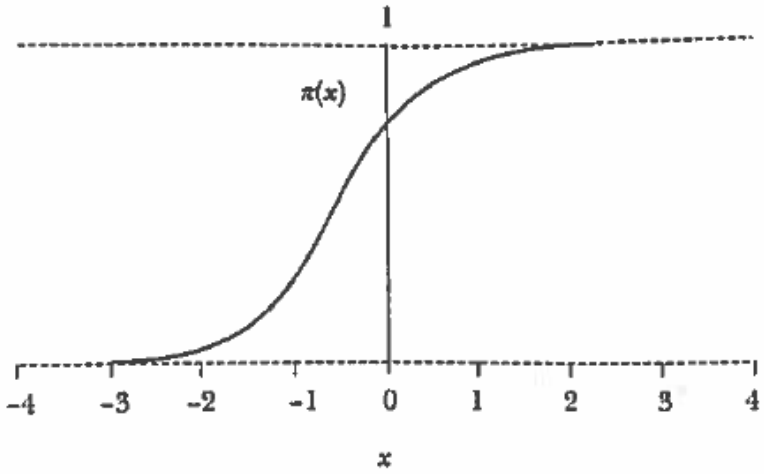
$$\frac{\pi_i}{1-\pi_i} = e^{\eta_i} = e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}} \qquad (9)$$

Po ďalších úpravách vzťahu sa dostaneme k vyjadreniu podmienenej strednej hodnoty π_i alternatívnej vysvetľovanej premennej pomocou nelineárnej funkcie k nezávisle premenných:

$$\pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}} = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}} \tag{17}$$

Grafom závislosti medzi podmienenou strednou hodnotou $\pi(x)$ a hodnotami nezávislej premennej X je pre prípad jednej vysvetľujúcej premennej ($k = 1$) logistická sigmoidná krivka, tzv. *S-krivka*. Jej vzhľad závisí od parametrov β_0, β_1 (Terek, Horníková, Labudová, 2010).

Graf č. 2 – Graf závislosti $\boldsymbol{\pi}$ od $\boldsymbol{x}$



Zdroj: Terek, Horníková, Labudová, 2010, s. 175

3.3.1 Šanca a pomer šancí

Výraz $\frac{\pi_i}{1-\pi_i}$, ktorý je súčasťou rovnice modelu logistickej regresie, sa nazýva **šanca** (*odds*). Vyjadruje podiel dvoch pravdepodobností, že udalosť nastane ($Y = 1$) a že udalosť nenastane ($Y = 0$). Pri interpretácii sa však využíva tzv. **pomer šancí**, ktorý sa označuje ako Odds Ratio (OR).

$$OR = \frac{odds_1}{odds_2} \tag{18}$$

kde $odds_1$ a $odds_2$ sú šance pre prvý a druhý s ním porovnávaný prípad.

3.3.2 Odhady parametrov modelu logistickej regresie

Na odhad parametrov modelu logistickej regresie sa používa metóda maximálnej vierohodnosti (*maximum likehood*). Pri odhade parametrov hľadáme maximum funkcie vierohodnosti, ktorá je vyjadrená tvarom:

$$L(p_i, \mathbf{y}) = \prod_{i=1}^{n} {p_i}^{y_i}(1 - p_i)^{1-y_i} \tag{19}$$

Pri hľadaní extrému – maxima, využívame logaritmickú funkciu vierohodnosti:

$$l(p_i, \mathbf{y}) = \ln\left[\prod_{i=1}^{n} {p_i}^{y_i}(1 - p_i)^{1-y_i}\right] = \sum_{i=1}^{n}[y_i \ln p_i + (1 - y_i)\ln(1 - p_i)] \tag{20}$$

Po odhade parametrov pomocou metódy maximálnej vierohodnosti nasleduje testovanie štatistickej významnosti modelu logistickej regresie a testovanie štatistickej významnosti jednotlivých parametrov.

3.3.3 Overenie štatistickej významnosti modelu logistickej regresie a jednotlivých parametrov

Test pomerom vierohodnosti

Pri tomto teste sa najskôr testuje štatistická významnosť modelu ako celku, čiže sa overuje, či platí nulová hypotéza H_0 (Terek, Horníková, Labudová, 2010):

H_0: Model nie je štatisticky významný (vektor regresných koeficientov je nulový), $\boldsymbol{\beta}_* = \mathbf{0}_\mathrm{k}$

H_1: Model je štatisticky významný (aspoň jeden regresný koeficient nie je nulový), $\boldsymbol{\beta}_* \neq \mathbf{0}_\mathrm{k}$

kde $\boldsymbol{\beta}_*$ je vektor regresných koeficientov β_j $(j = 1, 2, \dots, k)$ a $\mathbf{0}_\mathrm{k}$ je nulový vektor.

Pri teste vierohodnostným pomerom sa používa testovacia štatistika pomer vierohodností G:

$$G = -2\ln\left[\frac{L\ (model\ bez\ premennej)}{L(model\ s\ premennou)}\right] \tag{21}$$

Výraz v zátvorke sa nazýva pomer vierohodností (*likehood ratio*) a porovnáva funkciu vierohodnosti modelu bez premenných a funkciu vierohodnosti modelu so všetkými vstupnými premennými.

Ak platí nulová hypotéza, štatistika G má χ^2- rozdelenie s počtom stupňov voľnosti k, ktorý je daný rozdielom medzi počtom parametrov plného modelu a redukovaného modelu. V prípade modelu jednoduchej logistickej regresie je počet stupňov voľnosti 1. Kritická oblasť pre tento test je $W_\alpha = \{\chi^2;\ \chi^2 \geq \chi^2_\alpha(k)\}$.

Waldov test

Waldovým testom sa overuje zhoda vektora parametrov $\boldsymbol{\beta}$ a vektora známych konštánt $\boldsymbol{\beta}_0$. Overujeme platnosť nulovej hypotézy $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$ oproti alternatívnej hypotéze $H_1: \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. Na overenie sa používa Waldova testovacia štatistika (Šoltés, Vojtková, Šoltésová, 2018):

$$Wald = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)^{\mathrm{T}} . \mathbf{S}_{\mathbf{b}}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \tag{22}$$

$\widehat{\boldsymbol{\beta}}$ je vektor odhadov regresných koeficientov. $\mathbf{S}_{\mathbf{b}}^{-1}$ je variačno-kovariačná matica vektora odhadov regresných koeficientov $\widehat{\boldsymbol{\beta}}$. Waldova testovacia štatistika má asymptoticky χ^2-rozdelenie s k stupňami voľnosti (rovnajú sa počtu testovaných parametrov). Kritická oblasť pre tento test je $Wald = \{\chi^2;\ \chi^2 \geq \chi^2_\alpha(k)\}$.

Test štatistickej významnosti koeficienta $\boldsymbol{\beta_j}$ $(\boldsymbol{j = 0, 1, 2, \ldots k})$

Pri testovaní štatistickej významnosti parametrov modelu – koeficienta β_j overujeme platnosť nulovej hypotézy $H_0: \beta_j = 0$ – parameter β_j nie je štatisticky významný oproti alternatívnej hypotéze: $H_1: \beta_j \neq 0$ – parameter β_j je štatisticky významný. Ako testovaciu charakteristiku používame **Waldovu premennú**, ktorá má za predpokladu platnosti nulovej hypotézy tvar:

$$W(\beta_j) = \left(\frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}\right)^2 \tag{23}$$

$s_{\hat{\beta}_j}$ je odhad smerodajnej odchýlky bodového odhadu $\hat{\beta}_j$ parametra β_j. Waldova premenná má χ^2-rozdelenie s jedným stupňom voľnosti. $\frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$ má asymptoticky normované normálne rozdelenie $N(0,1)$. Kritická oblasť pre Waldovu premennú je $W = \{\chi^2;\ \chi^2 \geq \chi^2_\alpha(1)\}$.

Intervalové odhady pre parametre $\boldsymbol{\beta_j}$ modelu logistickej regresie

Na výpočet intervalových odhadov pre parametre β_j modelu logistickej regresie je možné použiť dve metódy. Prvá metóda používa pri výpočte funkciu vierohodnosti, ale výpočet je zložitý a náročný. Druhá predpokladá asymptoticky normálne rozdelenie bodových odhadov parametrov a vytvorené intervaly sa nazývajú Waldove intervaly.

Waldove $(1-\alpha)$ 100% intervaly spoľahlivosti pre lokujúcu konštantu β_0 a regresné koeficienty β_j vypočítame:

$$\hat{\beta}_0 - z_{1-\frac{\alpha}{2}}\, s_{\hat{\beta}_0} < \beta_0 < \hat{\beta}_0 + z_{1-\frac{\alpha}{2}}\, s_{\hat{\beta}_0} \tag{24}$$

$$\hat{\beta}_j - z_{1-\frac{\alpha}{2}}\, s_{\hat{\beta}_j} < \beta_j < \hat{\beta}_j + z_{1-\frac{\alpha}{2}}\, s_{\hat{\beta}_j} \tag{25}$$

$\hat{\beta}_0$ a $\hat{\beta}_j$ $(j = 1,2,\dots,k)$ predstavujú hodnoty bodových odhadov parametrov modelu, odhadnuté metódou maximálnej vierohodnosti. $s_{\hat{\beta}_0}$ a $s_{\hat{\beta}_j}$ sú odhady smerodajných odchýlok bodových odhadov $\hat{\beta}_0$ a $\hat{\beta}_j$ modelu a $z_{1-\frac{\alpha}{2}}$ je kvantil normovaného normálneho rozdelenia.

Intervalové odhady pomeru šancí

Nech je odhad pomeru šancí (OR) definovaný ako pomer šancí pre dva vektory hodnôt pozorovaní $\mathbf{x}_a$ a $\mathbf{x}_b$, pričom sa líšia v hodnotách spojitej premennej X_l o jednotku, tzn. $x_{al} - x_{bl} = 1$.

$(1-\alpha)100$ % interval spoľahlivosti pre pomer šancí má tvar:

$$exp\left(\hat{\beta}_l - z_{1-\frac{\alpha}{2}}\, s_{\hat{\beta}_l}\right) < OR < exp\left(\hat{\beta}_l + z_{1-\frac{\alpha}{2}}\, s_{\hat{\beta}_l}\right) \tag{26}$$

kde $\hat{\beta}_l$ predstavuje hodnotu bodového odhadu parametra $\hat{\beta}_l$, ktorý prislúcha premennej X_l a $s_{\hat{\beta}_l}$ je hodnota výberovej smerodajnej odchýlky tohto bodového odhadu $\hat{\beta}_l$.

3.3.4 Procedúry výberu premenných

Pri regresných modeloch sa využíva niekoľko procedúr výberu vstupných premenných, pričom cieľom je vybrať také premenné, ktoré budú štatisticky významne vplývať na vysvetľovanú premenné. V prípade, že máme veľký počet vysvetľujúcich

premenných vstupujúcich do modelu, naším cieľom je tento počet zredukovať, jednak z dôvodu náročnosti výpočtov (modely s veľkým počtom premenných sa náročne interpretujú), a zároveň, keďže nadhodnocujú rozptyly odhadov, môžu zvyšovať stupeň multikolinearity. Najznámejšie procedúry výberu premenných sú:

- metóda krokovej regresie (*forward* a *stepwise selection*),
- metóda postupnej eliminácie (*backward elimination*),
- metóda všetkých možných regresií (*all possible regressions*).

Metóda krokovej regresie vychádza z modelu, ktorý nemá žiadne vysvetľujúce premenné. Poznáme dva varianty krokovej regresie – starší variant (*forward selection*) a novší (*stepwise selection*). Metóda pracuje s každou vysvetľujúcou premennou tak, akoby mala byť zahrnutá v modeli a pre každú vypočíta jej prínos k vysvetleniu variability vysvetľovanej premennej. Kritériom, ktoré rozhoduje o tom, či bude premenná v danom kroku začlenená do modelu, je absolútna veľkosť parciálneho koeficienta korelácie medzi vysvetľovanou premennou a posudzovanou premennou pri eliminácii vplyvu premenných, ktoré boli do modelu zaradené v predchádzajúcich krokoch, veľkosť *F*-štatistiky pre test štatistickej významnosti prínosu posudzovanej vysvetľujúcej premennej k vysvetleniu variability cieľovej premennej alebo absolútna veľkosť *t*-štatistiky pre test štatistickej významnosti regresného koeficienta pri posudzovanej vysvetľujúcej premennej (bližšie (Šoltés, 2019)).

Prvá – staršia metóda krokovej regresie (*forward selection*) pracuje tak, že sa do modelu pridá tá premenná, ktorá je najviac korelovaná s vysvetľovanou premennou a postupne sa do modelu pridávajú ďalšie premenné, ktoré majú štatisticky významný prínos pre vysvetľovanú premennú na zvolenej hladine významnosti, až kým nezostanú tie vysvetľujúce premenné, ktoré štatisticky významný prínos nemajú. Premenné, ktoré sa pridali do modelu, v ňom zostávajú.

Druhá – novšia metóda krokovej regresie (*stepwise selection*) po každom vstupe premennej do modelu znovu otestuje štatistickú významnosť regresných koeficientov, ktoré stoja pri premenných už zaradených do modelu v predošlých krokoch. Pri tejto metóde nemusí byť zaradenie akejkoľvek vysvetľujúcej premennej definitívne, pretože môže byť v ktoromkoľvek nasledujúcom kroku z modelu vylúčená (Šoltés 2019, s. 106).

Metóda postupnej eliminácie (*backward elimination*) je opakom metódy krokovej regresie. Model, z ktorého vychádzame, je na začiatku tvorený všetkými vysvetľujúcimi

premennými. V ďalších krokoch sa z neho premenné postupne vylučujú, na základe ich štatistickej nevýznamnosti, až v modeli nezostanú len tie vysvetľujúce premenné, ktoré majú, na zvolenej hladine významnosti, štatisticky významný vplyv na vysvetľovanú premennú. Pri tomto postupne sa využíva test *F*-štatistiky alebo testy významnosti parciálnych koeficientov korelácie medzi vysvetľujúcou a vysvetľovanou premennou, pričom vplyv ostatných vysvetľujúcich premenných sa eliminuje (Šoltés 2019, str. 107).

Metóda všetkých možných regresií (*all possible regressions*) predstavuje vytvorenie regresných modelov, ktoré obsahujú všetky kombinácie vysvetľujúcich premenných. Ak sa v modeli nachádza *k* vysvetľujúcich premenných, odhadneme 2^k regresných modelov. Je niekoľko kritérií, ktoré sa následne využívajú na výber najlepšieho regresného modelu:

- **MSR – reziduálny rozptyl** je miera variability rezíduí, ktorá zohľadňuje počet vysvetľujúcich premenných aj rozsah súboru. Hodnota tohto rozptylu by mala byť čo najmenšia, pretože pri vstupe významnej vysvetľujúcej premennej do modelu hodnota charakteristiky klesne a pri nevýznamnej premennej vzrastie.
- $\boldsymbol{r^2_{adj}}$ **– upravený koeficient determinácie** (*adjusted coefficient of determination*) penalizuje model za počet vysvetľujúcich premenných. Najlepší model je ten, ktorý má upravený koeficient determinácie najväčší.
- **AIC – Akaikeho informačné kritérium** je charakteristika, ktorá by mala byť pri hľadaní najlepšieho modelu čo najmenšia.
- **BIC – Sawaovo bayesovské informačné kritérium** je kritérium, ktoré by rovnako ako AIC malo byť čo najmenšie pri hľadaní najlepšieho modelu.
- $\boldsymbol{C_p}$ **– Mallowsova štatistika** – v prípade vysokých hodnôt tejto štatistiky hovoríme o skreslení modelu.
- **SBC – Swarzovo bayesovské kritérium**
- **PC – Amemiyaovo predikčné kritérium**

Pre kritériá AIC, BIC, SBC a PC platí, že sa využívajú ako navzájom sa doplňujúce. Pri hľadaní najlepšieho modelu by mali mať tieto charakteristiky čo najnižšie hodnoty (Šoltés 2019).

3.4 Rozhodovacie stromy

Rozhodovacie stromy (*decision trees*) patria medzi neparametrické metódy učenia sa s učiteľom, ktoré sa využívajú na klasifikáciu a regresiu. Rozhodovací strom je štruktúra, ktorá sa používa na rozdelenie veľkého súboru prípadov v databáze na malé súbory prípadov postupnou aplikáciou jednoduchých rozhodovacích pravidiel (Terek, Horníková, Labudová, 2010). Podľa inej definície pozostáva rozhodovací strom zo súboru pravidiel, ktoré sa využívajú na rozdelenie veľkej heterogénnej populácie do menších homogénnych skupín, pričom sa rešpektuje výstupná premenná (Berry, Linoff, 2004).

Cieľom rozhodovacieho stromu je vytvoriť model, ktorý predikuje hodnotu cieľovej premennej naučením sa jednoduchých rozhodovacích pravidiel odvodených z dátových funkcií. V prípade, že cieľová (výstupná) premenná je kategoriálna, hovoríme o klasifikačných stromoch. Ak je cieľová (výstupná) premenná spojitá, hovoríme o regresných stromoch (Terek, Horníková, Labudová, 2010).

Rozhodovací strom sa zobrazuje graficky formou stromu a obsahuje koreňový uzol, listové a nelistové uzly a hrany. Koreňový uzol, ktorý sa tiež nazýva aj rozhodovací uzol, sa nachádza na nultej úrovni. Predstavuje výber, ktorého výsledkom bude rozdelenie všetkých záznamov na dve alebo viac vzájomne sa vylučujúcich množín. Uzly reprezentujú triedu alebo testovanú premennú. Uzol, ktorý sa ďalej nevetví, sa nazýva listový uzol, list alebo koncový uzol. Predstavuje konečný výsledok kombinácie rozhodnutí alebo udalostí. Uzly, ktoré sa ďalej vetvia, sa nazývajú nelistové alebo terminálové uzly. Horný okraj uzla je spojený s jeho nadradeným uzlom (buď priamo s koreňovým uzlom alebo s iným nelistovým uzlom) a spodný okraj je spojený s jeho podradenými uzlami (buď s ďalšími nelistovými uzlami alebo s listovými uzlami, pri ktorej sa daná vetva končí). Hrany reprezentujú hodnoty testovanej premennej.

Na rozdelenie nadradených uzlov na čistejšie podriadené uzly cieľovej premennej sa používajú iba vstupné premenné, ktoré súvisia s cieľovou premennou. Môžeme použiť diskrétne aj spojité vstupné premenné, či kategoriálne vstupné premenné. Pri zostavovaní modelu je nutné najskôr identifikovať najdôležitejšie vstupné premenné a následne rozdeliť v koreňovom uzle a v následných nelistových uzloch do dvoch alebo viacerých kategórií – podľa toho aké kategórie má vetviaca premenná. Testovacie kritériá sa vyberajú podľa typu výslednej premennej a budeme sa im venovať neskôr. Postup štiepenia rozhodovacieho stromu pokračuje až dovtedy, kým nie sú splnené vopred

stanovené kritériá homogenity alebo zastavenie. Vo väčšine prípadov sa na vytvorenie rozhodovacieho stromu nepoužijú všetky potenciálne vstupné premenné a v niektorých prípadoch sa môže dokonca použiť vstupná premenná aj viackrát na rôznych úrovniach rozhodovacieho stromu (Song a Lu, 2015).

Všeobecne pri zostavovaní štatistického modelu je potrebné zohľadňovať zložitosť a robustnosť daného modelu. Najmä pri rozhodovacích stromoch môže veľmi zložitý a široký model spôsobiť situáciu, že bude príliš dobre predpovedať hodnoty na trénovacej množine (všetky listové uzly budú 100% čisté, tzn. že všetky záznamy majú cieľový výsledok). Takýto rozhodovací strom by bol dokonale a úplne prispôsobený pozorovaniam v trénovacej množine a mal by v každom liste málo záznamov, takže by nemohol spoľahlivo predikovať budúce prípady, a teda by mal zlú zovšeobecniteľnosť (nebol by dostatočne robustný). Tomuto problému sa hovorí preučenie rozhodovacieho stromu (*overfitting*). Znamená to, že model síce dobre vysvetľuje vzťahy na trénovacej množine, ale tieto vzťahy nie sú všeobecne platné pre iné množiny údajov a dochádza k vysokej chybovosti. Aby sa zabránilo nadmernej zložitosti modelu a jeho preučeniu, musia sa pri vytváraní rozhodovacieho stromu aplikovať pravidlá zastavenia. Medzi tieto pravidlá patrí minimálny počet pozorovaní v liste, minimálny počet pozorovaní v uzle pred rozdelením alebo hĺbka rozhodovacieho stromu.

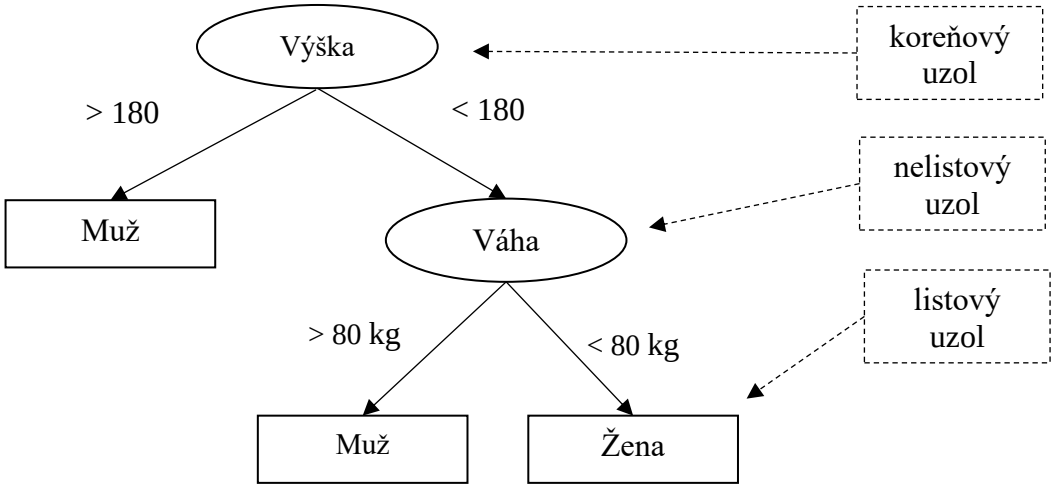
V niektorých situáciách spomínané pravidlá nefungujú dobre, a preto je alternatívnym spôsobom tzv. *orezávanie rozhodovacieho stromu*. Poznáme dva typy orezávania, a to: prepruning (orezávanie stromu počas jeho vytvárania) a postpruning (orezanie stromu až po jeho vytvorení). Pri metóde prepruning sa ukončí rast stromu predčasne prostredníctvom upraveného algoritmu. Používa sa chí-kvadrát test alebo metódy úpravy viacnásobným porovnávaním, aby sa zabránilo generovaniu nevýznamných vetiev. Pri metóde postpruning sa vygeneruje úplný rozhodovací strom a pri jeho orezávaní sa posudzuje, ako sa zhorší jeho klasifikačná schopnosť. Orezávanie znamená, že sa nelistové uzly zmenia na listové.

Ďalšou metódou výberu najlepšej alternatívy rozhodovacieho stromu, je využitie nielen trénovacej množiny, ale aj validačnej (tzn. že sa vzorka rozdelí na dve časti – jedna je trénovacia množina a druhá je validačná). Rozhodovací strom, ktorý je vytvorený pomocou trénovacej množiny je následne testovaný pomocou validačnej množiny (Song a Lu, 2015). Validačná množina slúži na výber podstromu z množiny všetkých potenciálnych stromov a následne je potom vytvorený strom nahradený tým

najvhodnejším podstromom. Ako kritérium sa používa miera nesprávnej klasifikácie (Terek, Horníková, Labudová, 2010).

Ukážeme si príklad binárneho rozhodovacieho stromu, ktorý bol vytvorený len na demonštratívne účely.

Obr. č. 2 – Graf rozhodovacieho stromu



Zdroj: Vlastné spracovanie

Vytvorený strom je dvojúrovňovou hierarchickou štruktúrou, ktorá sa skladá z piatich uzlov, a to z jedného koreňového uzla, troch listových uzlov a jedného nelistového uzla. Koreňový uzol a nelistový uzol predstavujú jednu vstupnú premennú a listové uzly obsahujú výstupnú premennú. Množina údajov obsahuje dve vstupné premenné – výšku v centimetroch a váhu v kilogramoch a výstupnú premennú pohlavie – muž alebo žena.

Rozhodovací strom je možné uložiť v grafickej podobe alebo ako súbor pravidiel. Napríklad pri uvedenom príklade klasifikačného rozhodovacieho stromu by bol súbor pravidiel nasledovný:

IF Výška > 180 cm THEN Muž

IF Výška <= 180 cm AND Váha > 80 kg THEN Muž

IF Výška <= 180 cm AND Váha <= 80 kg THEN Žena

3.4.1 Kritériá výberu testovacích premenných – klasifikačné rozhodovacie stromy

Na každej úrovni vetvenia sa v rozhodovacom strome vyberá testovacie kritérium. Dané kritérium závisí od toho, aký je dátový typ cieľovej premennej. Testovacie kritérium, ktoré je na danej úrovni zvolené, sa snaží zabezpečiť čo najvyššiu čistotu dcérskych uzlov, čo znamená, že sa snažia ponechať iba jednu triedu výstupnej premennej v danom uzle (Terek, Horníková, Labudová 2010).

Pri kategoriálnej výstupnej premennej sa ako testovacie kritérium používa entropia, Giniho index, informačný zisk alebo test nezávislosti. Pri spojitej výstupnej premennej môžeme buď vytvoriť jej kategórie a tým si zabezpečiť výber testovacieho kritéria pre kategoriálnu premennú, alebo výstupnú premennú ponecháme spojitú a v tom prípade ako testovacie kritérium použijeme redukciu rozptylu alebo F-test.

Entropia

Predpokladajme trénovaciu množinu s n prípadmi. Každý prípad opisuje hodnota vstupnej premennej A a hodnota výstupnej premennej Y. Vstupná premenná A nadobúda hodnoty $a_i (i = 1, 2, \dots, k)$ a výstupná premenná má m rôznych hodnôt $y_j (j = 1, 2, \dots, m)$. **Entropia** (*Entropy*, nazývaná aj miera nečistoty) je definovaná ako funkcia, ktorú pre prípad výstupnej premennej Y, vyjadríme v tvare:

$$H(Y) = -\sum_{j=1}^{m} \left(p_j \log_2 p_j\right) \tag{27}$$

kde p_j je pravdepodobnosť výskytu triedy j výstupnej premennej. Pravdepodobnosti p_j odhadujeme pomocou relatívnych početností $\frac{n_j}{n}$, kde n_j je absolútna početnosť triedy výstupnej premennej $y_j (j = 1, 2, \dots, m)$ v množine trénovacích prípadov (Berka, 2003).

Vzťah na výpočet entropie upravíme takto:

$$H(Y) = -\sum_{j=1}^{m} \left(\frac{n_j}{n} \log_2 \frac{n_j}{n}\right) \tag{28}$$

Predpokladajme, že výstupná premenná má iba dve kategórie. Ak všetky prípady patria do tej istej triedy, entropia nadobúda minimálnu hodnotu 0. Ak je početnosť tried výstupnej premennej rovnaká, entropia nadobúda maximálnu hodnotu 1.

Vstupná premenná A rozdelí trénovaciu množinu prípadov do k tried. Očakávaná entropia vstupnej premennej A, ktorá má k kategórií:

$$H(A) = \sum_{i=1}^{k} p_i H(a) = \sum_{i=1}^{k} \frac{n(a_i)}{n} H(a_i) \tag{29}$$

kde $n(a_i)$ je počet prípadov trénovacej množiny, ktoré nadobúdajú hodnotu a_i znaku A, $H(a_i)$ je hodnota entropie na množine prípadov, ktoré majú hodnotu a_i znaku A a N je počet prípadov v trénovacej množine.

$H(a_i)$ vypočítame nasledovne:

$$H(a_i) = -\sum_{j=1}^{m} \frac{n_j(a_i)}{n(a_i)} \log_2 \frac{n_j(a_i)}{n(a_i)} \tag{30}$$

kde $n_j(a_i)$ je počet prípadov trénovacej množiny z triedy j výstupného znaku, ktoré majú hodnotu a_i znaku A.

Ak je testovacím kritériom entropia, na vetvenie sa vyberá taká vstupná premenná, pre ktorú je hodnota očakávanej entropie premennej A najmenšia (Terek, Horníková, Labudová, 2010).

Informačný zisk

Informačný zisk (*Information gain*) a pomerný informačný zisk (*information gain ratio*) sú miery odvodené z entropie. Pre entropiu je definovaný informačný zisk $Z(A)$ premennej A. Informačný zisk je definovaný ako rozdiel entropie $H(Y)$ vyčíslenej pre celú množinu údajov a entropie $H(A)$ pre jednotlivé podmnožiny, ktoré vzniknú vetvením uzla podľa premennej A:

$$Z(A) = H(Y) - H(A) \tag{31}$$

Zatiaľ čo v prípade entropie sme hľadali premennú s jej najmenšou hodnotou, v prípade informačného zisku hľadáme premennú s najvyššou hodnotou. Uvedené kritérium má jednu nevýhodu – neberie do úvahy počet hodnôt vybranej premennej.

Dôležité je iba to, ako dobre táto premenná odlišuje od seba príklady rôznych tried. Preto sa niekedy používa ako kritérium pre voľbu premennej pomerný informačný zisk, ktorý berie do úvahy entropiu aj počet hodnôt premennej A, ktorá sa použila pri vetvení. Je definovaný ako podiel informačného zisku $Z(A)$ a vetvenia $V(A)$:

$$PZ(A) = \frac{Z(A)}{V(A)} \tag{32}$$

$$V(A) = -\sum_{i=1}^{k} \frac{n(a_i)}{n} \log_2 \frac{n(a_i)}{n} \tag{33}$$

Giniho index

Giniho index (*Gini index*) je definovaný nasledovne (Berka, 2003):

$$G(Y) = 1 - \sum_{j=1}^{m} p_j^2 \tag{34}$$

Pravdepodobnosti p_j (pravdepodobnosť výskytu triedy j výstupnej premennej) odhadujeme pomocou relatívnych početností $\frac{n_j}{n}$:

$$G(Y) = 1 - \sum_{j=1}^{m} \left(\frac{n_j}{n}\right)^2 \tag{35}$$

Maximálnu hodnotu nadobúda vtedy, ak sú kategórie vysvetľovanej premennej rovnako zastúpené v príslušnom uzle. Minimálnu hodnotu nadobúda, ak je v príslušnom uzle len jedna kategória vysvetľovanej premennej.

Očakávaná hodnota Giniho indexu pre premennú A:

$$G(A) = \sum_{i=1}^{k} \frac{n(a_i)}{n} G(a_i) \tag{36}$$

kde $G(a_i)$ je Giniho index na množine prípadov, ktoré majú hodnotu a_i premennej A.

$$G(a_i) = 1 - \sum_{j=1}^{m} \left(\frac{n_j(a_i)}{n(a_i)}\right)^2 \tag{37}$$

kde $n_j(a_i)$ je počet prípadov trénovacej množiny z triedy j výstupného znaku, ktoré majú hodnotu a_i premennej A.

Očakávaná redukcia nečistoty:

$$Z_G(A) = G(Y) - G(A) \tag{38}$$

Ak je testovacím kritériom Giniho index, na vetvenie sa vyberá taká vstupná premenná, pre ktorú je očakávaná hodnota Giniho indexu premennej A najväčšia.

Ako testovacie kritérium pre voľbu premennej pri klasifikačných stromoch je možné použiť aj test nezávislosti χ^2. Táto miera umožňuje vyhodnocovať vzájomnú závislosť medzi vstupnou premennou a výstupnou premennou Y – vyberie sa taká vstupná premenná A, pri ktorej je závislosť najvyššia (p-hodnota pri teste nezávislosti χ^2 je najmenšia).

3.4.2 Základné algoritmy rozhodovacích stromov

Historicky je známych veľa algoritmov, ktoré generujú rozhodovacie stromy, ako napríklad AID (*Automatic Iteraction Detection*), THAID (*Theta-Automatic Iteraction Detection*), ELISEE (*Exploration of Links and Interactions through Segmentation of an Experimental Ensemble*). Bližší opis týchto algoritmov uvádzajú vo svojej práci McArdle a Ritschard (2014).

Najznámejšie algoritmy rozhodovacích stromov sú a CHAID, ID3, C4.5 alebo CART.

CHAID (*CHi-squared Automatic Interaction Detection*) sa používa, ak je výstupná premenná nominálna, pričom vysvetľujúce premenné môžu byť nominálne aj ordinálne. Generuje nebinárny strom a na určenie počtu vetiev stromu a na delenie pozorovaní využíva χ^2-test (Terek, Horníková, Labudová, 2010).

Algoritmus ID3 sa považuje za veľmi jednoduchý algoritmus rozhodovacieho stromu. Vstupné premenné aj výstupná premenná sú kategoriálne a algoritmus pracuje ako klasifikátor. Vytvára rozhodovací strom pre dané dáta v štruktúre zhora nadol. V každom uzle stromu sa testuje jedna premenná a testovacím kritériom je buď entropia – vtedy hľadá minimálnu hodnotu kritéria alebo informačný zisk – vtedy hľadá maximálnu hodnotu kritéria (Fakir, Azalmad, Elaychi, 2020). Tento proces sa vykonáva

rekurzívne, kým nie sú hodnoty v každom uzle homogénne a uzol je listový (všetky hodnoty patria do jednej triedy).

Na použitie ID3 algoritmu musia byť splnené tieto podmienky:

- nekontradikčnosť – trénovacie prípady si navzájom neprotirečia,
- neredundantnosť – trénovací prípad sa v trénovacej množine nevyskytuje viacnásobne,
- vzájomná nezávislosť premenných, ktoré opisujú prípady.

C4.5 algoritmus vychádza z algoritmu ID3. Generuje rozhodovací strom pre danú množinu údajov rekurzívnym rozdelením záznamov. Vstupné premenné môžu byť nominálne, ordinálne aj kvantitatívne, výstupná premenná je kategoriálna. Algoritmus je neinkrementálny[9], buduje rozhodovací strom zhora nadol, ako testovacie kritérium sa využíva pomerný informačný zisk. Aj v tomto algoritme je nutné, aby bola splnená podmienka nekontradikčnosti.

CART (*Classification and Regression Trees*) algoritmus generuje binárne rozhodovacie stromy (Song a Lu, 2015). Vstupné údaje sú buď nominálne, ordinálne alebo kvantitatívne spojité. Výstupná premenná môže byť tiež nominálna, ordinálna alebo kvantitatívna spojitá. Ako testovacie kritérium sa používa Giniho index alebo entropia. V prípade kvantitatívnej závislej premennej je použité ako testovacie kritérium variabilita závislej premennej.

3.4.3 Výhody a nevýhody rozhodovacích stromov

Medzi výhody rozhodovacích stromov patria (Scikit-learn, 2020):

- sú jednoduché na pochopenie a interpretáciu, dajú sa vizualizovať,
- vyžadujú jednoduchú prípravu údajov. Iné techniky často vyžadujú napr. normalizáciu údajov, umelé premenné alebo odstránenie prázdnych hodnôt.
- dokážu spracovať číselné aj kategoriálne údaje,
- algoritmus dokáže narábať s veľkými a zložitými súbormi údajov,

[9] Inkrementálny algoritmus spracováva trénovacie prípady, ktoré prichádzajú do trénovacej množiny postupne a po príchode každého nového prípadu sa modifikuje (napr. sa upravia klasifikačné pravidlá). Neinkrementálny algoritmus nespracováva jeden trénovací prípad za druhým (postupne), ale spracováva celú množinu trénovacích prípadov naraz. Je vhodný na riešenie úloh, pri ktorých sú všetky trénovacie prípady z trénovacej množiny k dispozícii hneď na začiatku.

- používajú model tzv. bielej skrinky (*white box model*). Ak je daná situácia v modeli pozorovateľná, vysvetlenie situácie je jednoducho vysvetlené pomocou boolean logiky. Naopak, v modeli čiernej skrinky (*black box model,* napr. v umelej neurónovej sieti), môžu byť výsledky interpretované ťažšie.
- je možná validácia modelu pomocou štatistických testov. To umožňuje zohľadniť spoľahlivosť modelu.

Medzi nevýhody rozhodovacích stromov patria:

- tí, ktorí sa učia vytvárať rozhodovacie stromy, môžu vytvoriť príliš zložité stromy, ktoré dobre nezovšeobecňujú údaje. Tomu sa hovorí preučenie (*overfitting*). Tomuto problému sme sa venovali v úvodnej časti tejto kapitoly.
- náročnejšie spracovanie v prípade chýbajúcich údajov,
- neberie sa do úvahy vzájomná korelácia vstupných premenných,
- môžu byť nestabilné, pretože malé odchýlky v údajoch môžu mať za následok generovanie úplne iného stromu.

3.5 Random forest

Random forest alebo náhodný les je technika používaná pre klasifikáciu a regresiu a pozostáva z veľkého počtu rozhodovacích stromov, ktoré fungujú ako súbor. Náhodný les je nadstavbou nad rozhodovacími stromami a odstraňuje niektoré problémy, ktoré pri modeloch rozhodovacích stromov vznikajú (napríklad ich nestabilita). Náhodný les používa metódu bootstrap pri vytváraní rozhodovacích stromov a existujú dva spôsoby interpretácie týchto výsledkov. Bežnejší postup je založený na hlasovaní, zvyčajne v prípade klasifikácie a na priemere v prípade regresie. Prakticky to znamená, že každý individuálny rozhodovací strom vytvorí predpoveď danej kategórie a kategória s najvyšším počtom hlasov sa stane predpoveďou celého modelu.

Základný koncept náhodného lesa je jednoduchý, ale účinný – vychádza z predpokladu, že je lepšie využiť viacero menších modelov spolu (modelov rozhodovacích stromov) ako jeden celok (model náhodného lesa), ako iba jediný z týchto modelov. To znamená, že veľké množstvo relatívne nekorelovaných modelov (rozhodovacích stromov), ktoré fungujú ako celok, prekoná všetky jednotlivé modely, ktoré ho tvoria. Kľúčová je nízka korelácia medzi modelmi. Uvedieme to na príklade

portfólia, ktoré je tvorené investíciami s nízkou koreláciou (napríklad akcie a dlhopisy). Tak ako je toto portfólio väčšie ako súčet jeho častí, tak aj nekorelované modely môžu vytvárať predpovede súboru, ktoré sú presnejšie ako ktorákoľvek z jednotlivých predpovedí. Dôvodom, prečo sú rozhodovacie stromy ako celok lepšie je, že sa stromy chránia navzájom pred svojimi individuálnymi chybami (pokiaľ sa všetky nemýlia v rovnakej veci). Aj keď sa niektoré stromy môžu mýliť, veľa ďalších bude správnych, takže ako skupina sa rozhodovacie stromy môžu pohybovať správnym smerom (Yiu, 2019).

Náhodný les je súborová technika schopná vykonávať regresné aj klasifikačné úlohy s použitím viacnásobných rozhodovacích stromov a techniky nazývanej Bootstrap Aggregation, známej pod názvom „bagging“. Bagging je algoritmus strojového učenia určený na zlepšenie stability a presnosti algoritmov strojového učenia používaných pri štatistickej klasifikácii a regresii. Znižuje rozptyl a pomáha predchádzať preučeniu modelu (Arif, 2020). Najskôr sa vykoná náhodný výber s opakovaním z trénovacej množiny, pričom rozsah výberu je zvyčajne rovnaký ako rozsah trénovacej množiny. Následne sa vykoná trénovanie každého rozhodovacieho stromu na vytvorených náhodných výberoch dát s rovnakým rozdelením pravdepodobnosti pre všetky rozhodovacie stromy v lese (Nagpal, 2017).

Náhodný les sa skladá zo súboru rozhodovacích stromov $T_1, T_2, \dots, T_K$. Pre metódu náhodného lesa sa používajú binárne stromy typu CART – ide o tzv. klasifikačný a regresný rozhodovací strom. Ich hlavnou výhodou je, že sú jednoduché na pochopenie, interpretáciu aj grafické znázornenie. Dokážu pracovať so vstupnými údajmi, ktoré obsahujú chýbajúce hodnoty. Množina hodnôt sa rozdelí na dve časti – trénovaciu a validačnú množinu. Rozhodovací strom vznikne tak, že z trénovacej množiny Θ vyberieme náhodnú vzorku Θ_i veľkosti n. Trénovacie množiny pre jednotlivé rozhodovacie stromy T_i sú vybrané prostredníctvom náhodných výberov s opakovaním – ide o tzv. bootstrapové výbery. Keďže ide o výbery s opakovaním, tak je možné rozdeliť aj veľmi malé súbory na veľký počet trénovacích a validačných množín. Tieto výbery majú nevýhodu v tom, že nie sú nezávislé, niektoré pozorovania sú vybrané opakovane a niektoré naopak vôbec.

Pozorovania, ktoré sú v i-tom bootstrapovom výbere Θ_i sa použijú pri tvorbe stromu T_i – ide o trénovaciu množinu pre daný rozhodovací strom. Pozorovania, ktoré sa do tohto výberu nedostali sú potom použité ako validačná množina k odhadu jeho chyby.

Odhady chyby na testovacom súbore sa nazývajú *oob* odhady (out of bag, out of bootstrap sample). Celkový počet *oob* pozorovaní tvorí tretinu množiny hodnôt.

Regresnú alebo klasifikačnú funkciu rozhodovacích stromov možno vyjadriť nasledovne:

$$h(x, \Theta_1), h(x, \Theta_2), \dots, h(x, \Theta_K) \tag{39}$$

kde h je funkcia, x je prediktor a $\Theta_1, \Theta_2, \dots, \Theta_K$ sú nezávislé, rovnako rozdelené náhodné vektory (Breiman, 2001).

Pri využití modelu náhodného lesa pre klasifikáciu získame z každého rozhodovacieho stromu informáciu o zaradení každého pozorovania do výslednej kategórie. Výsledok klasifikácie modelu náhodného lesa je daný tzv. väčšinovým hlasovaním všetkých rozhodovacích stromov.

$$\hat{C}_{rf}^{K}(x) = v\text{äčš}inov\text{é}\_hlasovanie\left\{\hat{C}_i(x)\right\}_1^K \tag{10}$$

kde $\hat{C}_i(x)$ je výsledok klasifikácie i-teho stromu (Komprdová, 2012a).

V prípade, že je model náhodného lesa využitý pre regresiu, predikcia každého pozorovania je priemerom zo všetkých stromov:

$$\hat{f}_{rf}^{K}(x) = \frac{1}{K}\left(\sum_{i=1}^{K} T_i(x)\right) \tag{11}$$

kde $T_i(x)$ je hodnota pozorovania pri i-tom rozhodovacom strome T_i.

Náhodný les zvyšuje svoju presnosť tak, že sa vytvárajú veľké rozhodovacie stromy, ktoré sa neorezávajú. Cieľom je znížiť chybu celého lesa, nielen individuálneho stromu. Zároveň sa kombinovaním výsledkov jednotlivých stromov udržuje vhodná hodnota rozptylu (kombinácia je zabezpečená formou väčšinového hlasovania alebo priemerovania). Okrem toho je potrebné zabezpečiť aj nízku koreláciu medzi jednotlivými stromami. Ak vytvoríme výbery s opakovaním, ktoré nie sú navzájom nezávislé, budú výsledné stromy korelované, čo môže viesť k nadhodnoteniu výsledkov klasifikácie alebo predikcie. Zníženie korelácie medzi jednotlivými rozhodovacími stromami sa dosiahne náhodným výberom práve určitého počtu prediktorov. Pre každý strom sa najlepšie vetvenie pre daný uzol hľadá práve z m prediktorov $X_1, X_2, \dots, X_M$. To znamená, že náhodný les využíva nielen náhodný výber pozorovaní, ale aj náhodný výber prediktorov (Komprdová, 2012b).

3.5.1 Algoritmus tvorby náhodného lesa

Nech máme m prediktorov $X_1, X_2, \dots, X_M$, p vstupných premenných a množinu trénovacích údajov Θ.

1. Vytvoríme bootstrapový podsúbor Θ_i, ktorý má veľkosť n. Ide o trénovaciu množinu pre daný rozhodovací strom T_i.
2. Vyberieme náhodne m prediktorov $X_1, X_2, \dots, X_M$ z p vstupných premenných. Výber určitého počtu premenných (prediktorov) znižuje koreláciu medzi rozhodovacími stromami.
3. Na bootstrapovom súbore Θ_i vytvoríme strom T_i s použitím m náhodne vybraných premenných (hľadáme najlepšie rozdelenie daného uzla medzi vstupné premenné na dva dcérske uzly). Rast stromu sa zastaví vtedy, keď dosiahneme minimálne hodnoty veľkosti uzlov.
4. Zaradíme *oob* pozorovania (validačná množina) vytvoreným stromom a určíme výslednú klasifikačnú kategóriu alebo predikciu všetkých *oob* pozorovaní.

Kroky 1-4 sa opakujú do konečného počtu stromov v lese. Na záver zrátame celkový výsledok klasifikácie/predikcie celého lesa väčšinovým hlasovaním (pri klasifikácii) alebo priemerovaním (pri regresii).

Výber počtu m vstupných premenných pre náhodný výber a počtu n stromov v lese je na začiatku určený experimentálne a vyžaduje si určité skúsenosti. Parametre sa testujú tak, že spustíme model náhodného lesa niekoľkokrát s rôznymi nastaveniami parametrov a snažíme sa získať model s najmenšou celkovou chybovosťou. Po určitom čase začínajú stromy konvergovať k správnej hodnote *oob* odhadov. Minimálnu veľkosť lesa určíme ako počet stromov, kedy sa chyba *oob* odhadu s pribúdajúcim počtom stromov už nemení (Komprdová, 2012b).

Uvažujeme p vstupných premenných. Počet prediktorov, ktoré chceme v modeli využiť, určíme nasledovne:

- v prípade klasifikácie hodnota $m = \sqrt{p}$ a minimálna veľkosť uzla je jedna,
- v prípade regresie hodnota $m = \frac{p}{3}$ a minimálna hodnota listového uzla je päť.

V praxi však počet prediktorov závisí od konkrétneho problému a počet m je vhodné zvoliť podľa výsledkov testovania modelu s rôznymi nastaveniami. Cieľom je vybrať také m, kde je najmenšia chyba modelu.

Pri výbere prediktorov môže pomôcť aj meranie významnosti premenných, resp. aký vplyv na výstupnú premennú majú vstupné premenné. Poznáme dva typy merania významnosti premenných, a to:

- permutačná významnosť – metóda, ktorá je založená na permutáciách a na poklese klasifikačnej presnosti – využíva sa tu celková chyba R (*misclassification rate, error rate*) alebo
- princíp zníženia nečistoty – významnosť jednotlivých premenných zistená pomocou Giniho indexu (Berk, 2016).

Pri využití prvej metódy sú po vytvorení i-teho stromu na i-tom bootstrapovom súbore Θ_i *oob* pozorovania klasifikované pomocou stromu T_i do jedného z listových uzlov a je spočítaná presnosť ich klasifikácie (napr. percento správne klasifikovaných pozorovaní), resp. vypočíta sa ich predikčná chyba (*misclassification rate*), ktorú označíme v_k, *k* = 1, ..., *K*. Hodnoty m-tého prediktora z *oob* pozorovaní sú permutované (randomizované) a takto upravené pozorovania sú znova klasifikované i-tym rozhodovacím stromom T_i, pričom predikčná chyba *k*-teho stromu s randomizovanými hodnotami *m*-tého prediktora je označená ako v_{km}. Následne sa porovnáva presnosť m-tého prediktora bez permutácie a s permutáciou. Pre každý *m*-tý prediktor vypočítame priemer naprieč všetkými *K* stromami a permutačná významnosť *m*-tého prediktora je následne vyjadrená takto:

$$I_j = \frac{1}{K}\sum_{i=1}^{K}\left(v_{kj} - v_k\right) \qquad j = 1, \ldots, m \tag{12}$$

Ak je pokles výrazný, premenná X_i je významná. Ak sa však permutáciou nezmení presnosť klasifikácie, prediktor X_i je nevýznamný. Čím väčší je rozdiel medzi predikciami, tým je prediktor X_m významnejší (Berk, 2016).

Druhá metóda spočíta vo využití Giniho indexu, ktorý pre *m*-tý prediktor označíme ako $I_{GINI}(X_m)$. Zakaždým, keď sa nelistový uzol rozdelí podľa prediktora X_m, Giniho index pre dva dcérske uzly je menší ako rodičovský uzol. Pre *m*-tý prediktor zistíme hodnotu Giniho indexu tak, že zoberieme *k*-ty strom a sčítame hodnoty Giniho indexu vo všetkých uzloch, v ktorých je delenie tohto prediktora vybrané za najlepšie. Sčítanie všetkých poklesov Giniho indexu (*Gini reduction*) pre každú individuálnu premennú X_m naprieč všetkými stromami v náhodnom lese (počet stromov je označený ako *K*) nám dáva hodnotu dôležitosti danej premennej:

$$I_{GINI}(X_j) = \frac{1}{K}\sum_{k=1}^{K} I_{GINI}(X_j, k) \tag{13}$$

Čím je súčet vyšší, tým je prediktor X_m významnejší.

3.5.2 Výhody a nevýhody modelu náhodného lesa

Medzi výhody modelu náhodného lesa môžeme zaradiť nasledovné:

- model je možné použiť na klasifikáciu aj regresiu,
- model funguje dobre s kategoriálnymi aj spojitými premennými,
- metóda náhodného lesa je založená na algoritme bagging, pričom vytvára veľké množstvo rozhodovacích stromov a kombinuje výstupy všetkých stromov. Tým znižuje rozptyl, predchádza preučeniu modelu a tým zlepšuje stabilitu a presnosť modelu.
- dokáže automaticky spracovať chýbajúce hodnoty,
- nie je nutné žiadne škálovanie premenných – štandardizácia ani normovanie nie sú potrebné, pretože sa používa prístup založený na pravidlách namiesto výpočtu vzdialenosti,
- dokáže efektívne spracovať nelineárne parametre,
- model je robustný voči odľahlým pozorovaniam a dokáže ich spracovať,
- algoritmus modelu náhodného lesa je veľmi stabilný, aj keď sa do množiny údajov pridajú nové, celkový algoritmus tým nie je ovplyvnený, je ovplyvnený maximálne jeden z rozhodovacích stromov.

Nevýhody modelu náhodného lesa:

- zložitosť modelu – vytvára sa veľa stromov a tieto stromy sú navzájom kombinované. Napr. v programovacom jazyku Python, v knižnici sklearn sa vytvára 100 rozhodovacích stromov pri modeli náhodného lesa.
- dlhšia doba trénovania – model náhodného lesa potrebuje v porovnaní s modelom rozhodovacieho stromu oveľa viac času na trénovanie, pretože generuje veľa stromov (Kumar, 2019).

3.6 Miery využívané na hodnotenie modelov

Na hodnotenie vytvorených modelov, resp. na výber toho najlepšieho modelu je možné využiť rôzne druhy mier a kritérií. V našej práci budeme využívať klasifikačné miery, ROC krivku a *F*-skóre (Allison, 2012).

3.6.1 Klasifikačné miery

Hodnoty klasifikačných mier vychádzajú z tzv. matice zámen (*confusion matrix*), ktorá je vytvorená pre kategoriálnu vysvetľujúcu premennú s dvoma obmenami (Tab. č. 1). Na jej diagonále sa nachádzajú počty prípadov, pri ktorých model klasifikoval správne kategóriu vysvetľujúcej premennej. Mimo diagonály sú počty prípadov, pri ktorých model klasifikoval kategóriu vysvetľujúcej premennej nesprávne. Miery klasifikácie, ktoré budeme v práci využívať sú celková chyba *R* (*error rate, misclassification rate*), celková správnosť *A* (*accuracy*), presnosť *PR* (*precision*), senzitívnosť *SZ* (*sensitivity, recall*) a špecifičnosť *SP* (*specificity*).

$$Celková\ chyba\ R = \frac{E}{C} \tag{44}$$

kde *E* je počet nesprávne klasifikovaných prípadov a *C* je celkový počet prípadov validačnej množiny.

$$Celková\ správnosť\ A = \frac{C - E}{C} = 1 - R \tag{45}$$

Tab. č. 1 – Matica zámen pre dve kategórie vysvetľovanej premennej

Klasifikovaný modelom	**Skutočná trieda 1**	**Skutočná trieda 0**	**Spolu**
1	TP	FP	TP + FP
0	FN	TN	FN + TN
Spolu	TP + FN	FP + TN	TP + FP + FN + TN

Zdroj: Allison, 2012

TP (*true positive*) – počet prípadov správne zaradených do kategórie 1

FP (*false positive*) – počet prípadov nesprávne zaradených do kategórie 1

TN (*true negative*) – počet prípadov správne zaradených do kategórie 0

FN (*false negative*) – počet prípadov nesprávne zaradených do kategórie 0

Pomocou matice zámen môžeme celkový chybu R a celkovú správnosť A vypočítať nasledovne:

$$R = \frac{FP + FN}{TP + FP + TN + FN} \tag{46}$$

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{47}$$

Aj ostatné spomínané štatistiky možno vypočítať pomocou matice zámen. Presnosť vyjadruje správnosť pre triedu 1, čo znamená, že vyjadruje podiel správne predikovaných prípadov pre triedu 1 a všetkých prípadov triedy 1.

$$PR = \frac{TP}{TP + FP} \tag{48}$$

Senzitívnosť vyjadruje pomer správne klasifikovaných prípadov pre triedu 1 a celkového počtu pôvodných prípadov triedy 1.

$$SZ = \frac{TP}{TP + FN} \tag{49}$$

Špecifičnosť vyjadruje pomer správne klasifikovaných prípadov triedy 0 a celkového počtu pôvodných prípadov triedy 0. Niekedy sa využíva jej hodnota odrátaná od jednotky, čo vyjadruje pomer nesprávne klasifikovaných prípadov triedy 1 a všetkých pôvodných prípadov triedy 0 (Allison, 2012).

$$1 - SP = 1 - \frac{TN}{TN + FP} = \frac{FP}{FP + TN} \tag{14}$$

3.6.2 ROC krivka

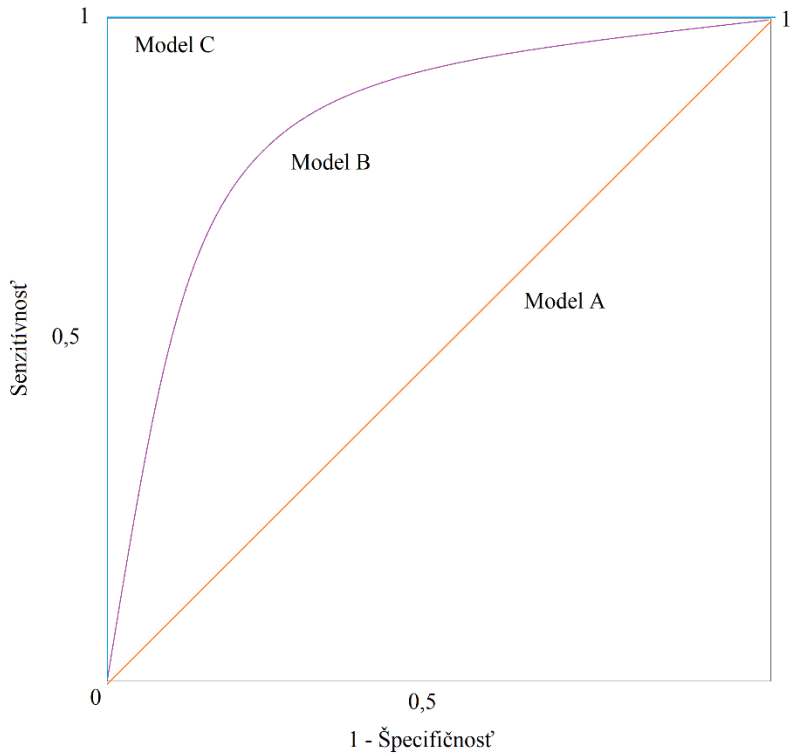
ROC krivka (*Receiver Operating Characteristic*) je nástroj, ktorý sa veľmi často využíva na porovnávanie prediktívnych modelov. ROC krivka, resp. ROC index vyjadruje, či prediktívny model správne klasifikuje *true positive* prípady (prípady správne zaradené do triedy 1) a *true negative* prípady (prípady správne zaradené do triedy 0). Kvalitný prediktívny model musí správne klasifikovať nielen prípady patriace do triedy 1, ale aj prípady patriace do triedy 0.

Na osi *y* sa nachádzajú hodnoty senzitívnosti SZ, ktoré vyjadrujú podiel prípadov správne zaradených do triedy 1 (*true positive*) a všetkých prípadov triedy 1. Na osi *x* sa

nachádzajú hodnoty 1 – špecifičnosť, ktoré vyjadrujú podiel prípadov správne zaradených do triedy 0 (*true negative*) a všetkých pôvodných prípadov triedy 0.

Na Obr. č. 3 sa nachádzajú ROC krivky troch prediktívnych modelov – A, B a C. Najlepší z modelov je model C, ktorý predstavuje ideálny prípad ROC krivky, kedy zaradí klasifikačné pravidlo každý prípad do správnej triedy s pravdepodobnosťou 1. Druhý najlepší model je model B a najhorší z porovnávaných modelov je model A. ROC krivka pre model A predstavuje náhodne rozdelenie prípadov do triedy 0 a 1 a je ROC krivkou pre základný (*baseline*) model (Reis, Patetta, 2021).

Obr. č. 3 – ROC krivka pre 3 prediktívne modely A, B a C



Zdroj: Reis, Patetta, 2021

3.6.3 F-skóre

F-skóre, tiež nazývané F_1-skóre, je mierou presnosti modelu. Používa sa na hodnotenie binárnych modelov, ktoré klasifikujú prípady do triedy 0 a triedy 1. Je kombináciou presnosti PR a senzitívnosti SZ modelu a je definované ako ich harmonický priemer. F_1-skóre vypočítame takto:

$$F_1 = \frac{2}{\frac{1}{SZ} + \frac{1}{PR}} = 2 * \frac{PR * SZ}{PR + SZ} = \frac{2 * TP}{2 * TP + FP + FN} \quad (15)$$

Čím je hodnota F_1-skóre vyššia, tým je model lepší. Toto skóre sa využíva na porovnanie prediktívnych modelov, ktoré sa snažia predpovedať rovnaký jav (Korstanje, 2021).

4 Výsledky práce

V tejto časti dizertačnej práce sa zameriame na praktickú aplikáciu uvedených metód na pripravených dátach. Vstupné údaje bude nutné najskôr upraviť, aby tvorili vhodný vstup na vytvorenie prediktívnych modelov. Konkrétne budeme vytvárať tri prediktívne modely: model binárnej logistickej regresie, model rozhodovacieho stromu a model náhodného lesa. Model, ktorý bude najlepšie predikovať vysvetľovanú premennú, je možné následne využiť pri cielení v online reklamnej kampani, kde cieľovou skupinou budú predplatitelia plateného obsahu spravodajskej stránky.

4.1 Zdroj dát

V našej práci sa zameriavame na predikciu predplatiteľov plateného obsahu stránky so spravodajstvom pochádzajúcej zo Spojeného kráľovstva (thetimes.co.uk). Okrem toho, že spoločnosť The Times vydáva aj tlačené noviny, disponujú aj online webovou stránkou so spravodajstvom, ktorá obsahuje množstvo článkov na rôzne témy.

Údaje pochádzajú z marketingovej agentúry GroupM Slovakia a s jej súhlasom boli poskytnuté za účelom vypracovania tejto dizertačnej práce.

Webová stránka, z ktorej dáta pochádzajú, neponúka žiadny obsah zadarmo, čiže ak chce návštevník čítať nejaký obsah článku, potrebuje sa zaregistrovať a získať dočasný prístup k vybraným článkom (tzv. trial verzia). Tento čas je ale 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 článku je motivované obsahom alebo typom článku, ktoré návštevníka zaujmú, keď navštevuje spravodajský web. Na toto rozhodnutie však môžu pôsobiť aj iné faktory. Dáta, ktoré máme k dispozícii, poskytujú behaviorálne informácie o používateľovi, pričom ide o informácie, ktoré sa dajú získať o akomkoľvek užívateľovi, ktorý vstúpi na internetovú webstránku. Cieľom dizertačnej práce je pomocou vytvoreného modelu predikovať, či internetový užívateľ patrí do cieľovej skupiny, konkrétne teda medzi návštevníkov webovej stránky, ktorí si predplatili platený obsah. Uvedená cieľová skupina je charakterizovaná určitými behaviorálnymi znakmi a pri snahe ju osloviť kdekoľvek na internete, bude na ňu cielená online reklama. Behaviorálne znaky jednotlivých užívateľov, resp. návštevníkov je identifikované z ich

online správania. Cieľom vlastníkov uvedenej spoločnosti so spravodajstvom je predať čo najväčšie množstvo online predplatného jednotlivým návštevníkom, ktorí prejavili záujem o danú stránku a to tým, že sa už zaregistrovali na uvedenej webstránke a získali dočasný prístup k článkom. Keďže chcú čo najviac z týchto dočasných prístupov prekonvertovať na platené prístupy, potrebujú takýchto návštevníkov (a rovnako tak návštevníkov im podobným, ktorí vykazujú podobné správanie na internet, a teda majú podobné behaviorálne charakteristiky) osloviť aj mimo uvedenej spravodajskej webstránky, a to konkrétne prostredníctvom online reklamy.

Dáta obsahujú informácie o 3808 užívateľoch, ktorí navštívili spravodajskú webstránku, z ktorých 1774 návštevníkov tvorilo práve takých, ktorí si po skončení dočasnej vezie predplatili platený obsah. Každý návštevník je v dátach uvedený iba raz. Zoznam premenných je uvedený v Tab. č. 2:

Tab. č. 2 – Zoznam premenných v dátovom súbore

Názov premennej	**Opis premennej**
USER_ID	Unikátne ID užívateľa
BROWSER	Typ prehliadača
DEVICE_TYPE	Typ využívaného zariadenia
GEO_DMA	ID Geo regiónu
OPERATING_SYSTEM	Verzia 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 The Times
HTTP_REFERER	Referenčná stránka pre prvú navštívenú stránku webu The Times
AUTOMOTIVE, BOOK_AND_LITERATURE, BUSINESS_AND_FINANCIAL, ...	Premenné klasifikujúce obsah navštívenej stránky pomocou taxonómie IAB[10] (spolu 30 premenných)
TARGET	Cieľová premenná – či užívateľ patrí medzi návštevníkov, ktorí si predplatili platený obsah webstránky (1) alebo nie (0)

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

[10] 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/.

Tab. č. 3 – Premenné klasifikujúce obsah web stránky

Názov premennej	**Preklad**
AUTOMOTIVE	Motorizmus
BOOKS_AND_LITERATURE	Knihy a literatúra
BUSINESS_AND_FINANCIAL	Podnikanie a financie
CAREERS	Kariéra
CONTENT_CHANNEL	Obsahový kanál
EDUCATION	Vzdelanie
EVENTS_AND_ATTRACTIONS	Udalosti a atrakcie
FAMILY_AND_RELATIONSHIPS	Rodina a vzťahy
FINE_ART	Výtvarné umenie
FOOD_AND_DRINK	Jedlo a pitie
HEALTHY_LIVING	Zdravý životný štýl
HOBBIES_AND_INTERESTS	Záľuby a koníčky
HOME_AND_GARDEN	Dom a záhrada
MEDICAL_HEALTH	Lekárstvo/ Zdravie
MOVIES	Filmy
MUSIC_AND_AUDIO	Hudba a zvuk
NEWS_AND_POLITICS	Správy a politika
PERSONAL_FINANCIAL	Osobné financie
PETS	Domáce zvieratá
POP_CULTURE	Popkultúra
REAL_ESTATE	Nehnuteľnosti
RELIGION_AND_SPIRITUALITY	Náboženstvo a spiritualita
SCIENCE	Veda
SHOPPING	Nakupovanie
SPORTS	Šport
STYLE_AND_FASHION	Štýl a móda
TECHNOLOGY_AND_COMPUTING	Technológia a výpočtová technika
TELEVISION	Televízia
TRAVEL	Cestovanie
VIDEO_GAMING	Videohry

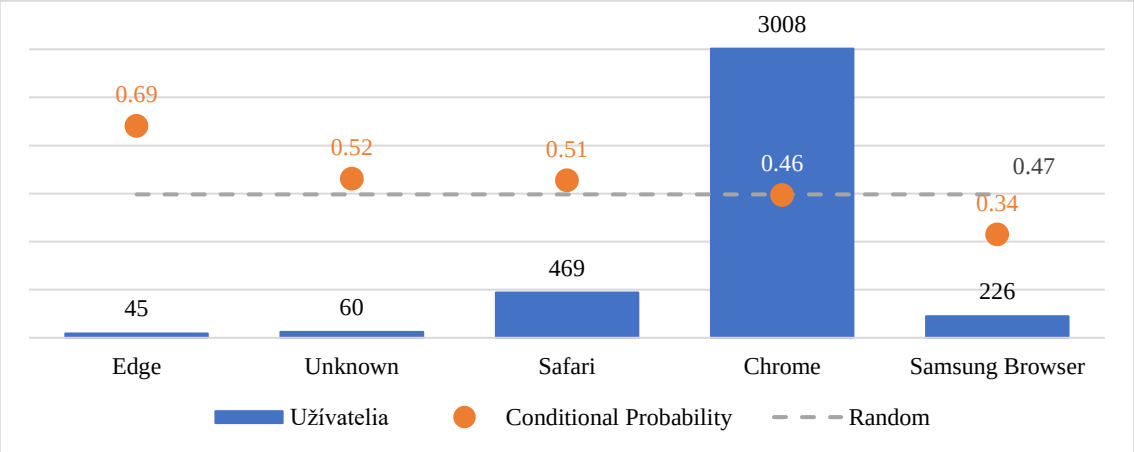
Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v Microsoft Excel

Niektoré z uvedených premenných budeme musieť pred vstupom do prediktívnych modelov upraviť. Ich opis a postup ich úpravy je uvedený v nasledujúcich podkapitolách. Ak budeme mať všetky premenné pripravené, rozdelíme množinu údajov na dve časti – trénovaciu a validačnú množinu v pomere 70:30, ktorý sa zvyčajne používa pri metódach učenia sa s učiteľom (Xu, Goodacre, 2018). Množinu údajov delíme z dôvodu otestovania presnosti modelov vytvorených na trénovacej množine. Toto výsledné testovanie prebieha na validačnej množine.

4.2 Preskúmanie vstupných premenných

V predchádzajúcej podkapitole sme predstavili vstupné premenné a teraz sa pozrieme na tieto premenné bližšie. Predpokladáme, že premenné bude nutné, pred vytváraním prediktívnych modelov, upraviť. Jednou z možností preskúmania vstupných premenných je ich vizualizácia, aby sme pochopili, ako sú premenné štruktúrované. Najskôr sa pozrieme na technologické a geografické premenné – *Browser, Device_type, Operating_system* a *Geo_DMA*. Graf č. 3 ukazuje rozdelenie premennej *Browser*. Okrem početností jednotlivých užívateľov podľa typu prehliadača, vidíme v grafe aj dve pravdepodobnosti – random a podmienenú pravdepodobnosť (*conditional probability*).

Graf č. 3 – Rozdelenie premennej *Browser* a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej



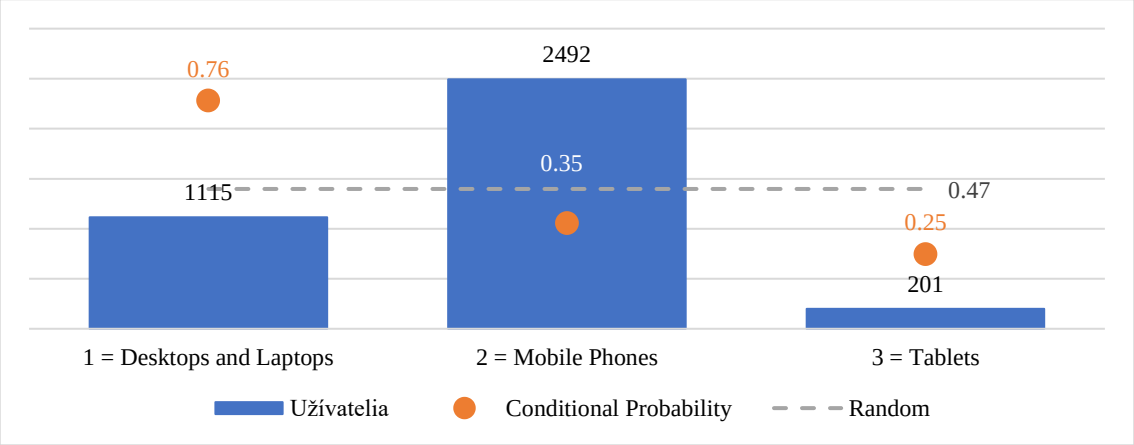
Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v Microsoft Excel

Random pravdepodobnosť nám hovorí o tom, že ak náhodne vyberieme užívateľa z dátového súboru, aká je pravdepodobnosť, že sa stane predplatiteľom plateného obsahu.

Táto pravdepodobnosť má hodnotu $p = 0,47$ a vypočítame ju ako podiel počtu užívateľov, ktorí si predplatili platený obsah a celkového počtu užívateľov (1774/3808 = 0,47).

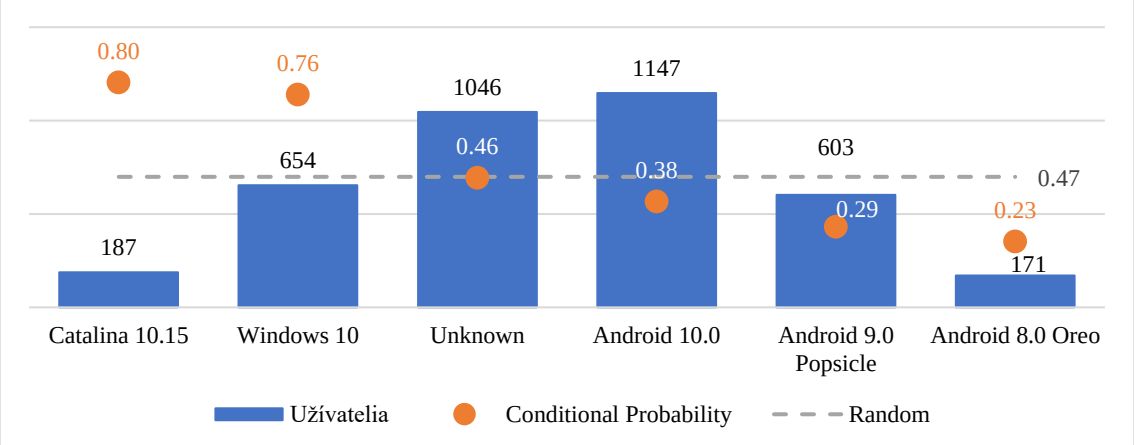
Podmienená pravdepodobnosť (*conditional probability*) za každý prehliadač je vypočítaná ako podiel užívateľov, ktorí využívajú daný typ prehliadača a zároveň si predplatili platený obsah a celkového počtu užívateľov v rámci daného typu prehliadača. Pozeráme sa na hodnoty tých kategórií danej premennej, ktoré sú väčšie ako random pravdepodobnosť, pretože práve pri týchto kategóriách je vyššia pravdepodobnosť, že si užívatelia predplatia platený obsah. Ak užívateľ používa prehliadač Edge, pravdepodobnosť, že si predplatí platený obsah je 69 %. Z toho vyplýva, že typy prehliadača, pri ktorých je najvyššia pravdepodobnosť, že si ich užívatelia predplatia platený obsah, využívajú browser Edge, Safari alebo neznámy prehliadač (Unknown).

Graf č. 4 – Rozdelenie premennej *Device_type* a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej



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

Graf č. 5 – Rozdelenie premennej *Operating_system* a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej

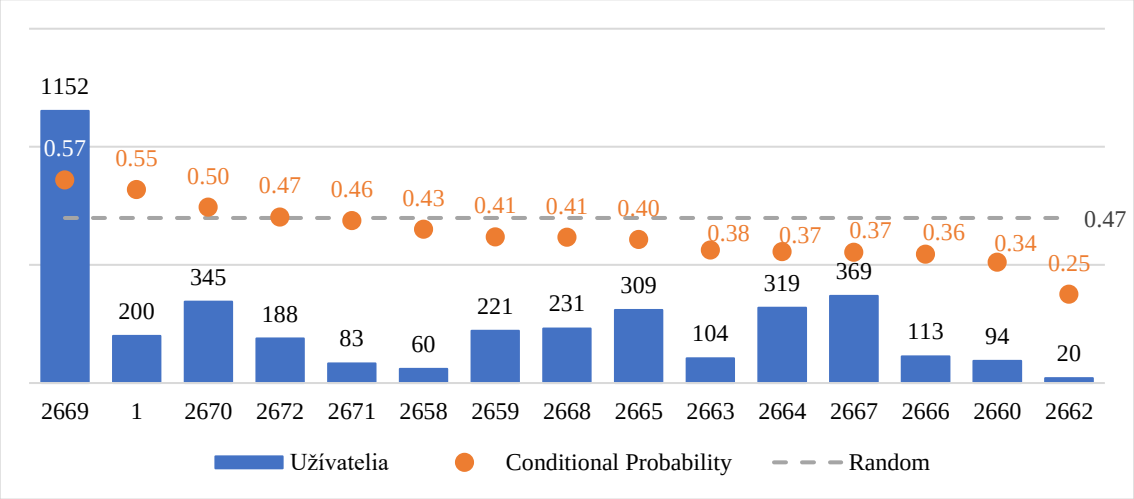


Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v Microsoft Excel

Podľa grafu č. 4 je najvyššia pravdepodobnosť, že sa užívateľ stane predplatiteľom plateného obsahu, v prípade, že využíva desktop alebo laptop. Táto pravdepodobnosť je až 76 %. Naopak najmenšia pravdepodobnosť predplatenia si plateného obsahu je u osôb, ktoré využívajú tablet, a to 25 %.

Najvyššiu pravdepodobnosť na predplatenie plateného obsahu na webovej stránke majú podľa Grafu č. 5 užívatelia využívajúci operačný systém Catalina 10.15 (macOS) a Windows 10.

Graf č. 6 – Rozdelenie premennej *Geo_DMA* a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej



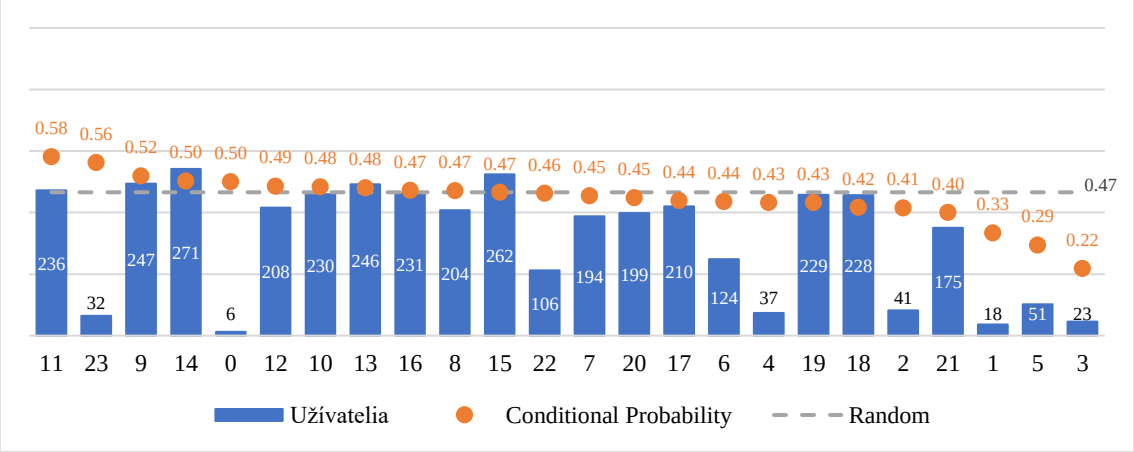
Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v Microsoft Excel

Z premennej *Geo_DMA* vidíme, že najvyššia pravdepodobnosť, že sa užívateľ stane predplatiteľom, je pri hodnotách 2669 a 1. Ostatné hodnoty sú pod úrovňou random pravdepodobnosti. Pri tejto premennej máme k dispozícii len kódy jednotlivých regiónov a nie ich presné názvy. Premenná *Geo_DMA* má menší vplyv na pravdepodobnosť predplatenia si obsahu ako predchádzajúce dve premenné *Device_type* a *Operating_system*, v prípade ktorých boli medzi príslušnými kategóriami väčšie disparity v predmetnej pravdepodobnosti, ako tomu je pri premennej *Geo_DMA* (podmienené pravdepodobnosti z rozpätia 0,25 a 0,57).

Ako ďalšie sme preskúmali premenné *Hour_of_day* a *Day_of_week*. Grafy č. 7 a č. 8 nás okrem iného informujú, že najvyššia pravdepodobnosť, že sa užívatelia stanú predplatiteľmi plateného obsahu je, ak navštívili webovú stránku o 11., 23. alebo 9. hodine ráno, v utorok, v nedeľu alebo v pondelok.

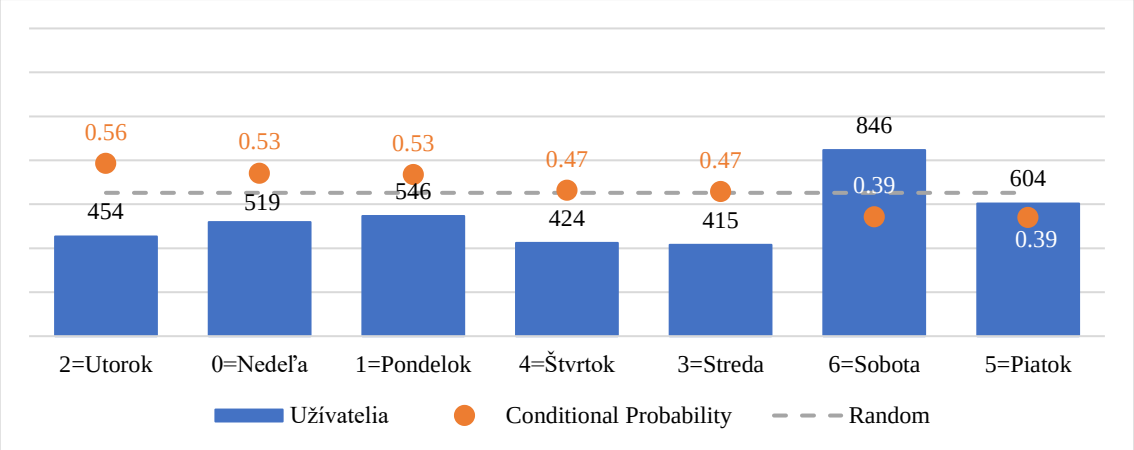
Z premenných *Hour_of_day* a *Day_of_week* vytvoríme nové premenné, ktoré budú opisovať afinitu[11] cieľovej skupiny voči hodine počas dňa a dňa v týždni. Tzn., že vytvoríme premennú *Affinity_HoD* a *Affinity_DoW*, ktoré budú numerické.

Graf č. 7 – Rozdelenie premennej *Hour_of_day* a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej



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

Graf č. 8 – Rozdelenie premennej *Day_of_week* (Nedeľa=0) a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej



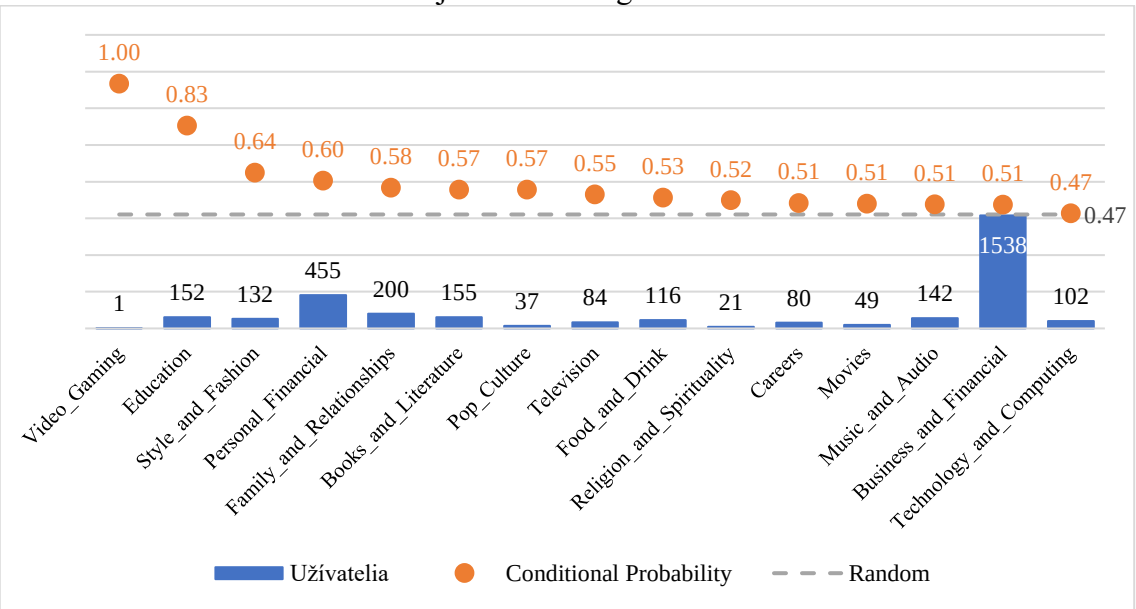
Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v Microsoft Excel

Pre premenné, ktoré tvoria IAB kategórie, sme kvôli prehľadnosti vytvorili dva nasledujúce grafy (Graf č. 9 a č. 10). Tieto premenné sú binárne – obmena 0 znamená, že obsah článkov, ktoré užívateľ navštevuje, nepatrí do danej kategórie a obmena 1, že do

[11] 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.

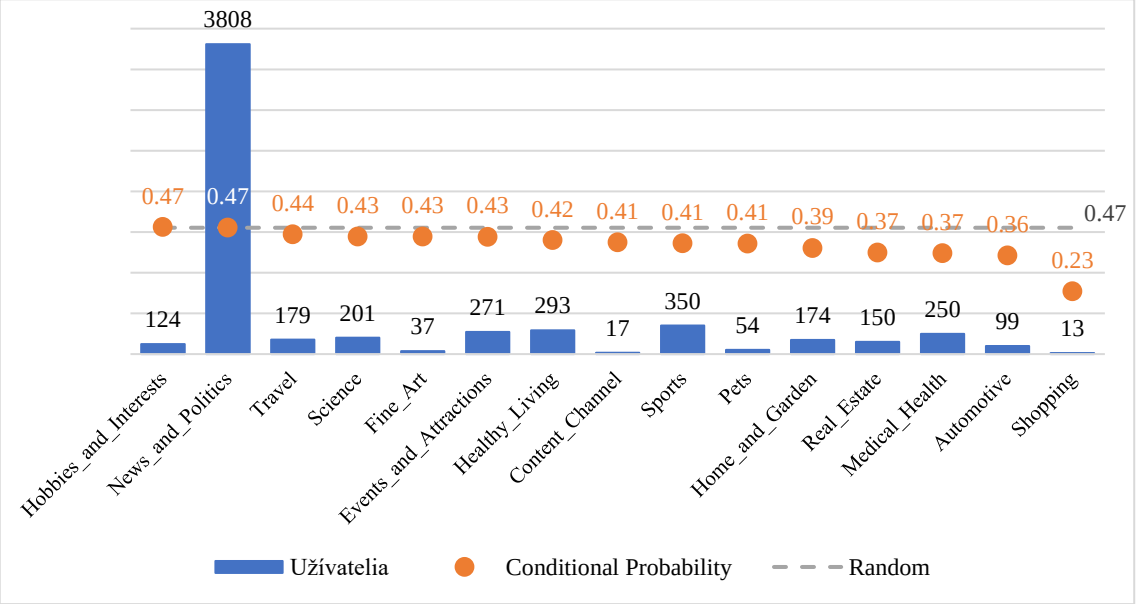
kategórie patrí (napr. premenná *Education*=1 znamená, že užívateľom navštevovaný obsah článkov patrí do kategórie *Education* – čiže do kategórie Vzdelávanie).

Graf č. 9 – Premenné charakterizujúce IAB kategórie – 1. časť



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

Graf č. 10 – Premenné charakterizujúce IAB kategórie – 2. časť



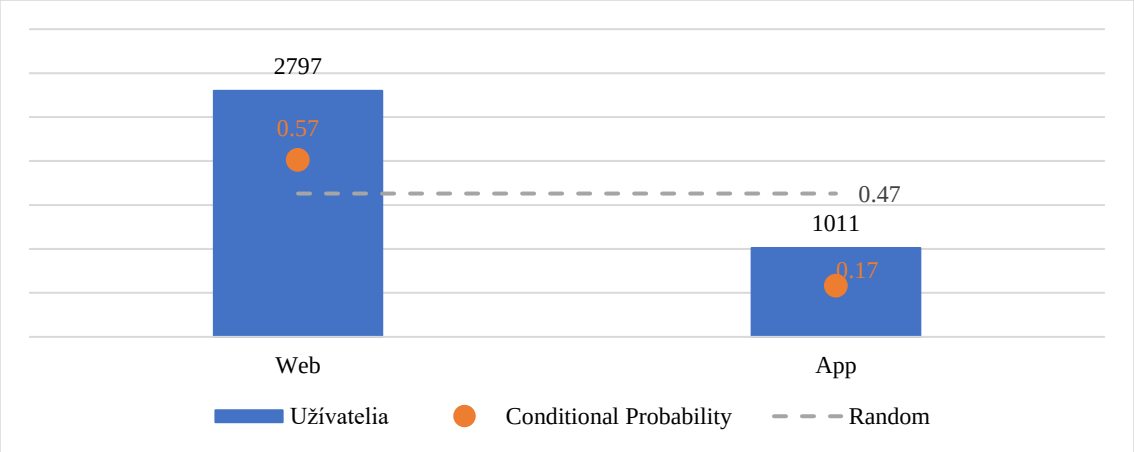
Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v Microsoft Excel

Z premenných, ktoré opisujú IAB kategórie, patria medzi tie s najväčšou pravdepodobnosťou predplatenia si obsahu, premenné *Video_Gaming*, *Education*, *Style_and_Fashion*, *Personal_Financial*, *Family_and_Relationship*,

Books_and_Literature, *Pop_Culture*, *Television*, *Food_and_Drink*, *Religion_and_Spirituality*, *Careers*, *Movies*, *Music_and_Audio* a *Business_and_Financial*.

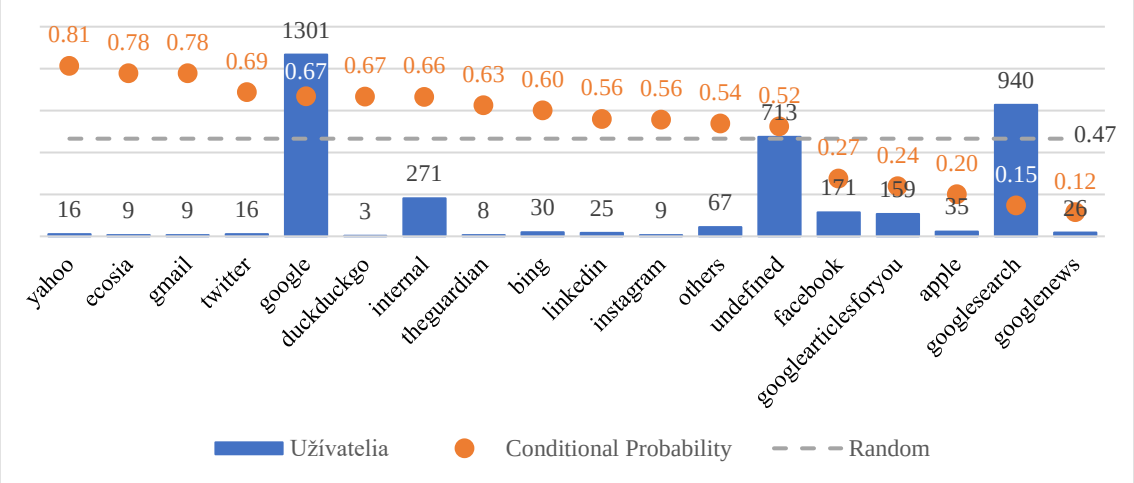
Následne sa zameriame na URL (*Uniform Resource Locator*) konkrétnych stránok a referenčných stránok. Referenčné stránky nás informujú o tom, z akého zdroja na danú web stránku užívateľ prišiel. Z pôvodnej premennej *Http_referer* sme vytvorili dve premenné – najskôr prvú premennú *Source_type*, ktorá 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.

Graf č. 11 – Rozdelenie premennej *Source_type* a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej



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

Graf č. 12 – Rozdelenie premennej *Source_ platform* a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej



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

Vyššiu pravdepodobnosť stať sa predplatiteľmi majú užívatelia, ktorí prišli z webovej stránky, napr. yahoo, ecosia alebo cez google vyhľadávač, z twitteru alebo gmailu.

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

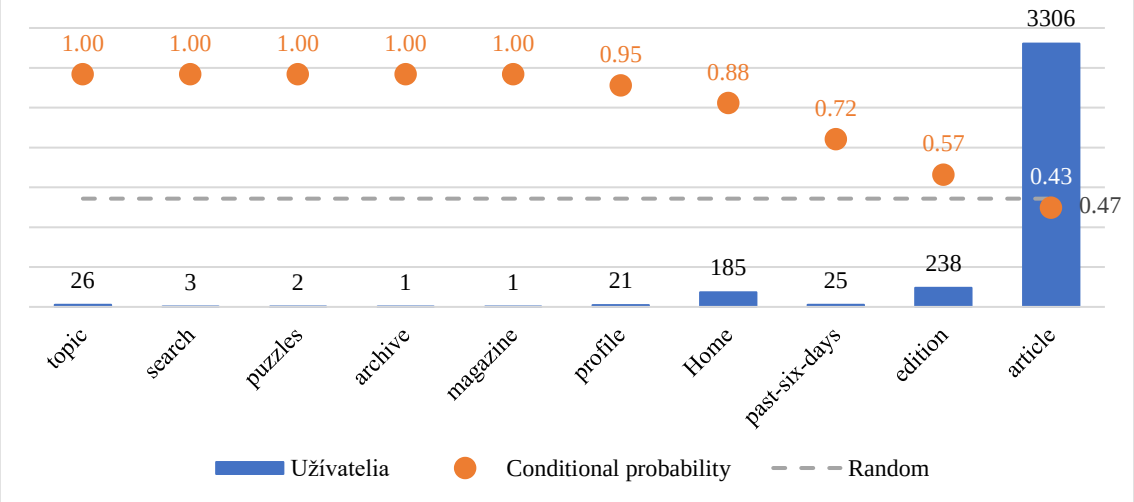
https://www.thetimes.co.uk/article/paradise-lost-the-latest-from-the- ... - hf6mvjhps

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*. Tieto kategórie zobrazíme v Grafe č. 13 a 14:

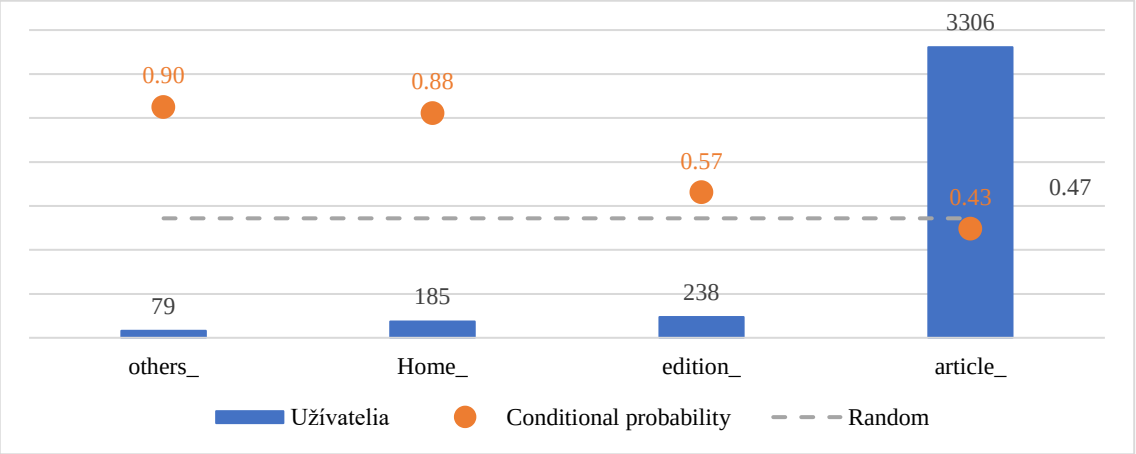
Graf č. 13 – Rozdelenie premennej *Sekcia_1* a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej



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

Podmienená pravdepodobnosť v Grafe č. 14 je vyrátaná ako podiel užívateľov, ktorí si predplatili platený obsah v rámci jednotlivých sekcií stránky a všetkých užívateľov v danej sekcii. Vidíme, že najpočetnejšou kategóriou premennej *Sekcia_1_a_2* je kategória article_ – teda patrí medzi konkrétne články.

Graf č. 14 – Rozdelenie premennej *Sekcia_1_a_2* a pravdepodobnosti predplatenia si plateného obsahu webstránky pre jednotlivé kategórie tejto premennej



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

Zvyšná časť premennej *URL*, ktorú sme označili ako *článok*, bude predmetom analýzy v nasledujúcej podkapitole. Keďže ide o názov článku zložený z viacerých slov, na analýzu bude použitá metóda Text Miningu.

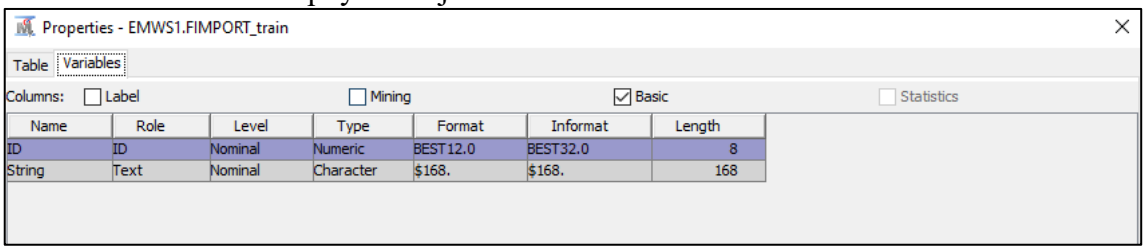
4.3 Text Mining a úprava URL

4.3.1 Načítanie vstupných údajov do SAS Enterprise Miner

Niektoré vstupné premenné z dátového súboru budeme musieť pred vstupom do prediktívnych modelov upraviť. V predchádzajúcej podkapitole sme už upravili niektoré pôvodné premenné a vytvorili z nich nové (napr. *Source_type*, *Source_platform*). Keď sme sa bližšie pozreli na hodnoty premennej *URL* (ktorá obsahuje webovú adresu článku, ktorý užívateľ čítal), zistili sme, že sa skladá z niekoľkých častí – okrem hlavnej webstránky, resp. domény obsahuje informácie, ktoré sme označili ako *sekcia_1*, *sekcia_2* a *článok*. Zo *sekcie_1* a *sekcie_2* sme vytvorili nové premenné a v tejto časti budeme analyzovať časť *článok* pomocou Text Miningu. Na túto analýzu bude využitý softvér SAS Text Miner (SAS TM), ktorý je súčasťou softvéru SAS Enterprise Miner (SAS EM). Najskôr vytvoríme csv súbor s názvom Data_url, do ktorého vložíme identifikátor (premennú *User_ID*) a časť *článok* (premenná *String*) z premennej *Url* tak, že jednotlivé spojovníky nahradíme medzerami. Predtým, ako môžeme spracovať dáta prostredníctvom SAS Text Minera, musíme vytvoriť projekt, knižnicu a diagram.

Po načítaní dátového súboru môžeme vstupné údaje skontrolovať a pozrieť sa na vlastnosti vstupných premenných: názov, rola, úroveň, typ, formát a dĺžka. V našom súbore sú iba dve premenné – *User_ID* a *String*. Vstupné údaje a teda všetky jednotlivé výrazy, ktoré plánujeme v načítanom dokumente spracovávať, nazývame korpus.

Obr. č. 4– Tabuľka vstupných údajov a ich vlastnosti



Name	Role	Level	Type	Format	Informat	Length
ID	ID	Nominal	Numeric	BEST12.0	BEST32.0	8
String	Text	Nominal	Character	$168.	$168.	168

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

Obr. č. 5 – Náhľad na vstupný súbor Data_url

EMWS1.FIMPORT_train

Obs #	ID	String
1	1	paradise lost the latest from the seychelles maldives mauritius and sri lanka hf6mvjhps
2	2	jeremy clarkson cheat love bray let me put my ass on the line and tell you that the donkey sex scene was real 2zrpzczz0
3	3	defence chiefs face battleover plan to scrap tanks ws87tdgbg
4	4	michel barnier refuses to open talks on brexit fishing deal wnffc56xv
5	5	danny cowley i shook the huddersfield chairman s hand and told him i d prove him wrong 278xk2thh
6	6	revealed viktor fedotov is tycoon behind aquind energy project pq0868vmj

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

4.3.2 Syntaktická analýza

Syntaktická analýza je zabezpečená prostredníctvom uzla Text Parsing. Tento uzol začína rozdelením textu jednotlivých pozorovaní vstupného súboru na tzv. tokeny. Identifikovaný zoznam tokenov (slov) možno upravovať v rôznych krokoch, podľa nastavenia vlastností. Tieto nastavenia výrazne ovplyvňujú počet výrazov, ktoré sa použijú na analýzu v uzloch.

Text mining sa zaoberá najmä spoločným výskytom pojmov v dátovom súbore. Spoločný výskyt sa v lingvistike chápe ako nadštandardná frekvencia výskytu dvoch termínov z textového súboru vedľa seba v určitom poradí. Spoločný výskyt možno interpretovať ako indikátor sémantickej blízkosti alebo idiomatický výraz. Hľadanie koreňa slova (*stemming*) a synonymá zohrávajú kľúčovú úlohu v identifikácii vzťahov medzi jednotlivými slovami v dokumente alebo viacerých dokumentoch. Napríklad, ak budeme vnímať slová „letecké spoločnosti“, „letectvo“ a „lietadlo“ ako podobné pri analýze názvov jednotlivých dokumentov, budeme vedieť, že tieto dokumenty spolu

nejako súvisia. Funkcia *stemming* v SAS TM umožňuje automaticky priraďovať synonymá k výrazom, ktoré môžu byť reprezentované ich koreňovými tvarmi slov, napríklad slovo *help* je koreňovým výrazom pre slová *help, helps, helping* a *helped*. Podobne sme slová spájajúce sa s pandémiou covid-19 (*covid, covid-19, coronavirus*) spojili pod jedno koreňové slovo, a to *coronavirus*.

Niekedy sa môže stať, že funkcia spojí slová, ktoré majú síce rovnaký koreň slova, ale sú významovo odlišné. Preto je nutné po aplikácii *stemmingu* na textový dokument bližšie preskúmanie týchto výsledkov. Algoritmus, ktorý vykonáva hľadanie koreňa slova, funguje na základe pravidiel alebo slovníka. Napríklad jedno z pravidiel je, že ak anglické slovo končí na „ed“, tak sa táto prípona zo slova odstráni.

Okrem synoným, ktoré sú definované pomocou synonymického slovníka, sme zadefinovali aj vlastné zoznamy synoným. Ich príklad vidíme na Obr. č. 6:

Obr. č. 6 – Tabuľka synoným

Synonyms-EMWS1.TextParsing_synonymDS

Child Term	Parent Term	Term Role	Parent Role
sas	sas institute		COMPANY
boris	boris johnson		PERSON
johnsons	boris johnsons		PERSON
jeremy clarkson	jeremy clarkson		PERSON
jeremy corbyns	jeremy corbyns		PERSON
enola holmes	enola holmes		PERSON
covid 19	covid 19		NOUN
can sharon	sharon white		PERSON
rishi sunak	rishi sunak		NOUN GROUP

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

Každý výraz má priradenú rolu, ktorá predstavuje buď slovný druh výrazu, klasifikáciu entity výrazu alebo hodnotu s názvom „noun group“, čo predstavuje skupinu zvyčajne dvoch výrazov, ktoré k sebe patria. Nastavená rola hovorí o tom, ako sa použijú uvedené synonymá, keď je povolené alebo zamietnuté používanie konkrétnych slovných druhov alebo entít.

Pri syntaktickej analýze je potrebné nastaviť aj jazyk, v ktorom v ktorom majú byť vstupné údaje spracované (v našom prípade sme tento jazyk nastavili na angličtinu, keďže dáta, ktoré máme k dispozícii, sú v angličtine).

Ďalším krokom bola identifikácia tokenov z hľadiska slovných druhov – s využitím algoritmov POS (*Part-of-Speech*). Túto funkcionalitu, ktorú poskytuje softvér môžeme a nemusíme využiť, ak sa rozhodneme, že slovný druh ku slovu priradíme sami. Niekedy nemá zmysel túto funkciu využívať, napríklad v komentároch na internete, v textových správach alebo ak sa využívajú neúplné vety. Rozhodli sme sa toto nastavenie ponechať.

Okrem algoritmov POS vie SAS TM identifikovať aj skupiny výrazov, ktoré k sebe patria – tzv. skupina podstatných mien (*Noun group*). Napríklad slová „data“ a „model“ sú analyzované ako jednotlivé výrazy a sú tiež identifikované ako skupina – „data model“.

Po identifikácii slovných druhov bol nasledujúcim krokom vynechanie niektorých slovných druhov z Text Mining analýzy. Rozhodli sme sa vynechať nasledovné slovné druhy: pomocné slovesá, spojky, citoslovcia, príčastia, predložky a zámená. Vynechanie konkrétnych slovných druhov je jednou z metód filtrovania výrazov. Ďalšou metódou, bez toho, aby sa vynechávali konkrétne slovné druhy, je používanie tzv. *štart* a *stop* zoznamov – štart zoznam predstavuje zoznam výrazov, ktoré chceme do analýzy zaradiť a stop zoznam predstavuje zoznam výrazov, ktoré chceme z analýzy vynechať.

Pri Text Mining analýze sa využíva aj funkcionalita slúžiaca na detekciu entít. Označenie entity je jednou z dôležitých informácií, ktorá je obsiahnutá v textových údajoch. SAS Text Miner dokáže identifikovať rôzne typy entít, napr. názov ulice, názov spoločnosti, organizácia, poštová adresa, meno osoby, telefónne číslo, e-mailová adresa, dátum, názov mesta a pod.. V našej práci sme napríklad identifikovali entity ako *Boris Johnson, Jeremy Clarkson, Jeremy Corbyns*. Tieto entity predstavovali mená osôb. Ďalšie entity, ktoré boli identifikované predstavovali napríklad organizáciu (*Scottish government, Britist Army*), časové obdobie (*14 next years, four years*), mieru (*25m, 32m, 27m*) atď.

Keďže vstupné údaje predstavujú veľké množstvo výrazov, je otázkou, či sú všetky tieto výrazy dôležité. Napriek tomu, že sme postupne odstránili niektoré slovné druhy, stále je veľa výrazov, ktoré nemusia byť pre analýzu veľmi prínosné. Preto sa jednotlivým výrazom priraďujú dva typy váh – lokálne a globálne. Hodnoty týchto váh sme nastavili nasledovne: pre lokálne – *Log* a pre globálne – *Entropy*. Na Obr. č. 7 vidíme 10 výrazov s najvyššou hodnotou entropie. Najvyššiu hodnotu entropie majú práve výrazy, ktoré majú nízku frekvenciu, ale sú pre analýzu dôležité.

Obr. č. 7 – Najdôležitejšie výrazy z hľadiska entropie

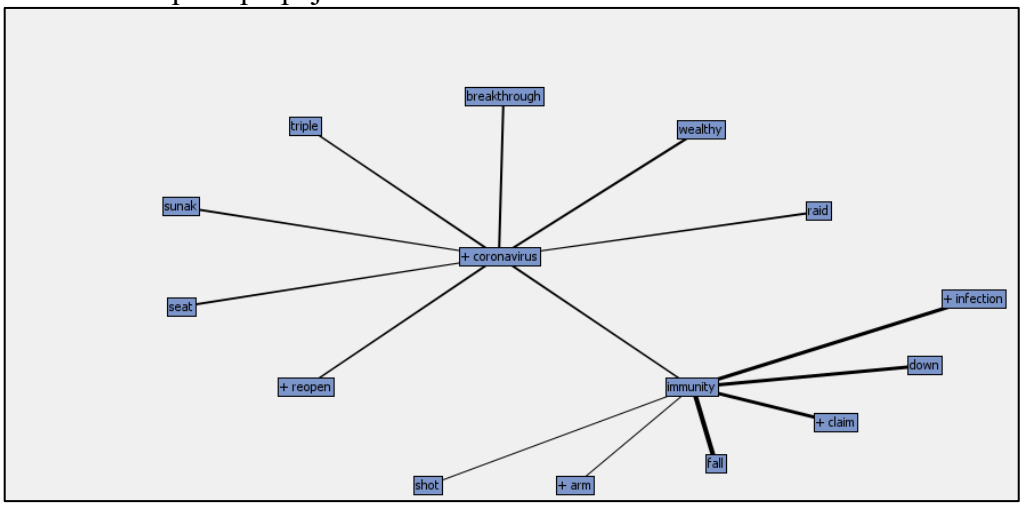
Terms

	TERM	FREQ	# DOCS	KEEP	WEIGHT ▼	ROLE	ATTRIBUTE
⊞	investment	6	4	☑	0.837	Noun	Alpha
	strike	5	4	☑	0.837	Noun	Alpha
⊞	walk	5	4	☑	0.837	Noun	Alpha
⊞	dad	5	4	☑	0.837	Noun	Alpha
	carbon	4	4	☑	0.83	Adj	Alpha
	low	4	4	☑	0.83	Adv	Alpha
	brian	4	4	☑	0.83	Noun	Alpha
	upad	4	4	☑	0.83	Noun	Alpha
	air	4	4	☑	0.83	Noun	Alpha
	chieftan	4	4	☑	0.83	Noun	Alpha

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

Vzťahy medzi jednotlivými výrazmi je možné vizualizovať pomocou koncepčných prepojení (Obr. č. 8). Koncepčné prepojenia pomáhajú pochopiť vzťahy medzi slovami na základe spoločného výskytu v dokumente. Výrazy v texte môžeme preskúmať výberom kľúčového slova (napr. *coronavirus*) a skúmaním slov, ktoré sú s týmto kľúčovým slovom spojené. Sila asociácie jednotlivých slov je znázornená hrúbkou čiary, ktorá jednotlivé slová spája. Čím je čiara hrubšia, tým je prepojenie silnejšie. Napr. výraz „*coronavirus*“ je najviac spojený s výrazmi „*breakthrough*“ (pokrok), *wealthy* (bohatý), *triple* (trojnásobný) alebo *immunity* (imunita, odolnosť).

Obr. č. 8 – Koncepčné prepojenia



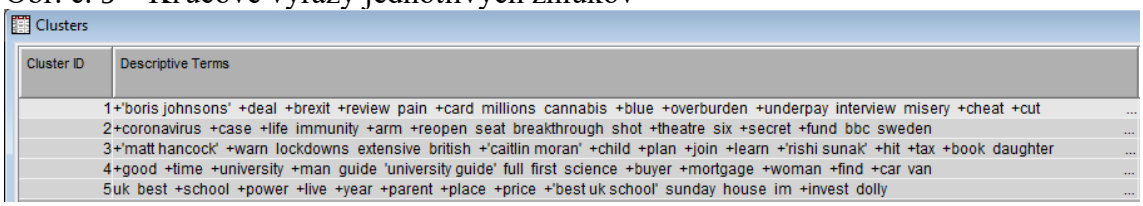
Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v SAS TM

4.3.3 Vytvorenie zhlukov a špecifikovanie tém z jednotlivých výrazov

Po syntaktickej analýze, preskúmaní a výbere jednotlivých dôležitých výrazov nasleduje vytvorenie zhlukov z týchto výrazov. V SAS TM sú dostupné dva typy algoritmov pre zhlukovanie – hierarchický algoritmus (*hierarchical algorithm*) a algoritmus maximalizácie očakávaní (*expectation-maximization (EM) algorithm*). EM zhlukovanie automaticky vyberá medzi dvoma verziami algoritmu – štandardným a škálovaným. Štandardná verzia EM algoritmu analyzuje všetky údaje v kroku iterácie a používa sa, keď máme malý dátový súbor. Škálovaná verzia EM algoritmu využíva časť vstupných údajov v každej iterácii a používa sa, keď máme veľký dátový súbor.

Pomocou zhlukovej analýzy sme vytvorili 5 zhlukov. Zhluky sú definované pomocou výrazov vybraných v procese filtrovania, ktoré boli vstupom do zhlukovej analýzy. Zhluk 1, ktorý tvorí 16 % výrazov, je reprezentovaný pojmami ako „Boris Johnson“, „deal“, „Brexit“, „review“, „card“, „million“, „overburden“, „underpay“, „misery“ a pod. Zhluk 1 spája teda články, ktoré sa týkajú politiky. Zhluk 2 tvorí 24 % výrazov, napr. „coronavirus“, „life“, „immunity“, „reopen“, „breakthrough“ a pod. Tento zhluk spája články, ktoré sa týkajú pandémie covid-19. Zhluk 3 je tvorený 13 % výrazov – „Matt Hancock“, „warn“, „lockdowns“, „Caitlin Moran“, „plan“, „join“, „learn“. Zhluk 3 spája články, ktoré sa týkajú rád a odporúčaní (či už od konkrétnych osobností alebo od expertov). Zhluk 4 tvorí 25 % výrazov, konkrétne „good“, „time“, „university“, „university guide“, „first“, „full“, „science“ a pod. Tento zhluk spája články týkajúce sa vzdelania a univerzít. Posledný je Zhluk 5 a ten tvorí 23 % výrazov – „uk“, „best“, „school“, „power“, „live“, „year“, „parent“, „place“, „best uk school“, „price“, „house“ a pod. Tento zhluk spája články s tematikou školstva a bývania.

Obr. č. 9 – Kľúčové výrazy jednotlivých zhlukov



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

Pomocou zhlukovej analýzy sme vytvorili 5 nových binárnych premenných, každú za jeden zhluk (*Cluster_**, hodnota 1 vyjadruje, že výrazy patria do uvedeného

zhluku a hodnota 0 vyjadruje, že do tohto zhluku nepatria). Tieto budú ďalej využité ako vstupné premenné pri vytváraní prediktívnych modelov.

4.4 Prediktívne modely

Naším cieľom bolo identifikovanie profilu internetových užívateľov, ktorí patria do cieľovej skupiny prostredníctvom vytvorenia prediktívnych modelov. Cieľovú skupinu tvoria návštevníci, ktorí si predplatili platený obsah na spravodajskej stránke. Prvý model, ktorý sme vytvorili, bol model binárnej logistickej regresie, druhý bol model rozhodovacieho stromu a tretí model náhodného lesa (*random forest*).

Medzi premenné, ktoré slúžili na vytvorenie modelov, sme zaradili aj tie premenné, ktoré sme získali pomocou Text Mining analýzy v predchádzajúcej podkapitole. Pred vytvorením modelov sme rozdelili dátový súbor na dve časti – trénovaciu množinu a validačnú množinu v pomere 70:30. Keďže naša vysvetľovaná premenná bola kategoriálna, metóda výberu bola nastavená na stratifikovaný náhodný výber.

4.4.1 Model logistickej regresie

Ako prvý z prediktívnych modelov sme vytvorili 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.

Vysvetľujúce premenné, ktoré sme pri tvorbe modelu použili, sme uviedli v úvode tejto kapitoly. Väčšiu časť premenných tvorili kategoriálne premenné, numerickými premennými boli *Affinity_HoD* a *Affinity_DoW*.

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*,

- premenná *Geo_DMA* mala referenčnú kategóriu *1*[12],
- 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*.

Pred vytvorením modelu logistickej regresie sme upravili niekoľko nastavení:

- väzbovú funkciu sme nastavili na funkciu *Logit*,
- nastavili sme vytvorenie modelu s lokujúcou konštantnou,
- spôsob vytvárania umelých premenných sme nastavili na *GLM* kódovanie (kódovanie 1, 0).

Po vytvorení modelu binárnej logistickej regresie sme dostali test pre významnosť modelu ako celku (Obr. č. 10), testy pre významnosť jednotlivých parametrov (Obr. č. 11) a odhad týchto parametrov a pomerov šancí (Tab. č. 4).

Najskôr sa pozrieme na overenie štatistickej významnosti modelu ako celku a štatistickú významnosť jednotlivých parametrov. Hypotézy, ktoré sme testovali:

H_0: *Model logistickej regresie nie je ako celok štatisticky významný*

H_1: *Model logistickej regresie je ako celok štatisticky významný*

Obr. č. 10 – 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 – thetimes.co.uk website; Vlastné spracovanie v SAS EM

[12] Premenná *Geo_DMA* má jednotlivé regióny označené iba kódom, k bližším informáciám sme sa nedostali

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 na vysvetľovanú premennú.

Obr. č. 11 – Test významnosti jednotlivých premenných

Type 3 Analysis of Effects

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	1	4.0835	0.0433
DeviceType	2	49.0223	<.0001
Education	1	38.8753	<.0001
GeoDMA_V2	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	3	44.5005	<.0001

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

Tab. č. 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[13]	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

[13] DeviceType obmeny: 1 = dektop a laptop, 2 = mobilný telefón, 3 = tablet

Parameter		Beta	p-value	Odds Ratio
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[14]	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

[14] G/GS = Google/Googlesearch

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

Model logistickej regresie mal okrem lokujúcej konštanty 47 nenulových parametrov. Kategoriálne premenné boli v modeli nahradené umelými premennými (*dummy variables*), ktorých počet je vždy o 1 nižší, ako je počet kategórií danej premennej.

Keďže 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 nezmenené.

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 používajú desktop alebo laptop 3,612-násobne vyššia ako u tých, ktorí používajú 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, šanca predplatenia plateného obsahu bola pri prekliku cez platformu Google je:

- 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 vplyv na vysvetľovanú premennú práve päť klasifikačných premenných. Šanca, že si používateľ predplatí platený obsah je:

- 6,834-násobne vyššia, ak čítaný článok patril do kategórie Vzdelanie (*Education*) oproti tomu, ak článok nebol v danej kategórii,
- 2,977-násobne vyššia, ak patril článok do kategórie Zdravý životný štýl (*Healthy Living*) oproti tomu, ak článok nebol v danej kategórii,
- 2,519-násobne vyššia, ak patril článok do kategórie Štýl a móda (*Style and Fashion*) oproti tomu, ak článok nebol v danej kategórii,
- 1,997-násobne vyššia, ak patril do kategórie Knihy a literatúra (*Books and Literature*) oproti tomu, ak článok nebol v danej kategórii a
- 2,151-násobne nižšia, ak patril do kategórie Lekárstvo/ Zdravie (*Medical Health*) oproti tomu, ak článok nebol v danej kategórii.

Z Text Mining analýzy vyšiel štatisticky významný iba zhluk č. 2, ktorý spájal články s tematikou pandémie covid-19. 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.

Predikované hodnoty vysvetľovanej premennej vychádzajú z predikovanej podmienenej pravdepodobnosti vypočítanej podľa hodnôt vysvetľujúcich premenných (vstupujúcich do modelu) pri danom pozorovaní. S využitím odhadov parametrov logitového modelu uvedených v Tab. č. 4 sme vypočítali pravdepodobnosť, že si užívateľ predplatí platený obsah na spravodajskej webstránke (Tab. č. 5).

Tab. č. 5 – Hodnoty vysvetľujúcich premenných 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 – desktop/laptop**	1	0
DeviceType	**2 – mobilný telefón**	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

Predikovaná podmienená pravdepodobnosť			
Parameter	**Kategórie**	**Pravdepodobnosť $\hat{p}$**	
		Užívateľ - typické hodnoty	**Užívateľ - netypické hodnoty**
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

Na výpočet bol využitý vektor vysvetľujúcich premenných, kde sme najskôr využili typické hodnoty pre cieľovú skupinu – internetoví užívatelia, ktorí si s najväčšou pravdepodobnosťou predplatia platený obsah na spravodajskej webstránke a pri druhom výpočte sme využili hodnoty vysvetľujúcich premenných, ktoré nie sú typické pre danú cieľovú skupinu. Hodnoty vidíme v Tab. č. 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.

V prípade, že hodnota niektorej premennej nebude tvoriť uvedenú typickú hodnotu, podmienená pravdepodobnosť bude o niečo nižšia, ale bude stále dostatočne vysoká na to, aby model dokázal predikovať cieľovú skupinu efektívne.

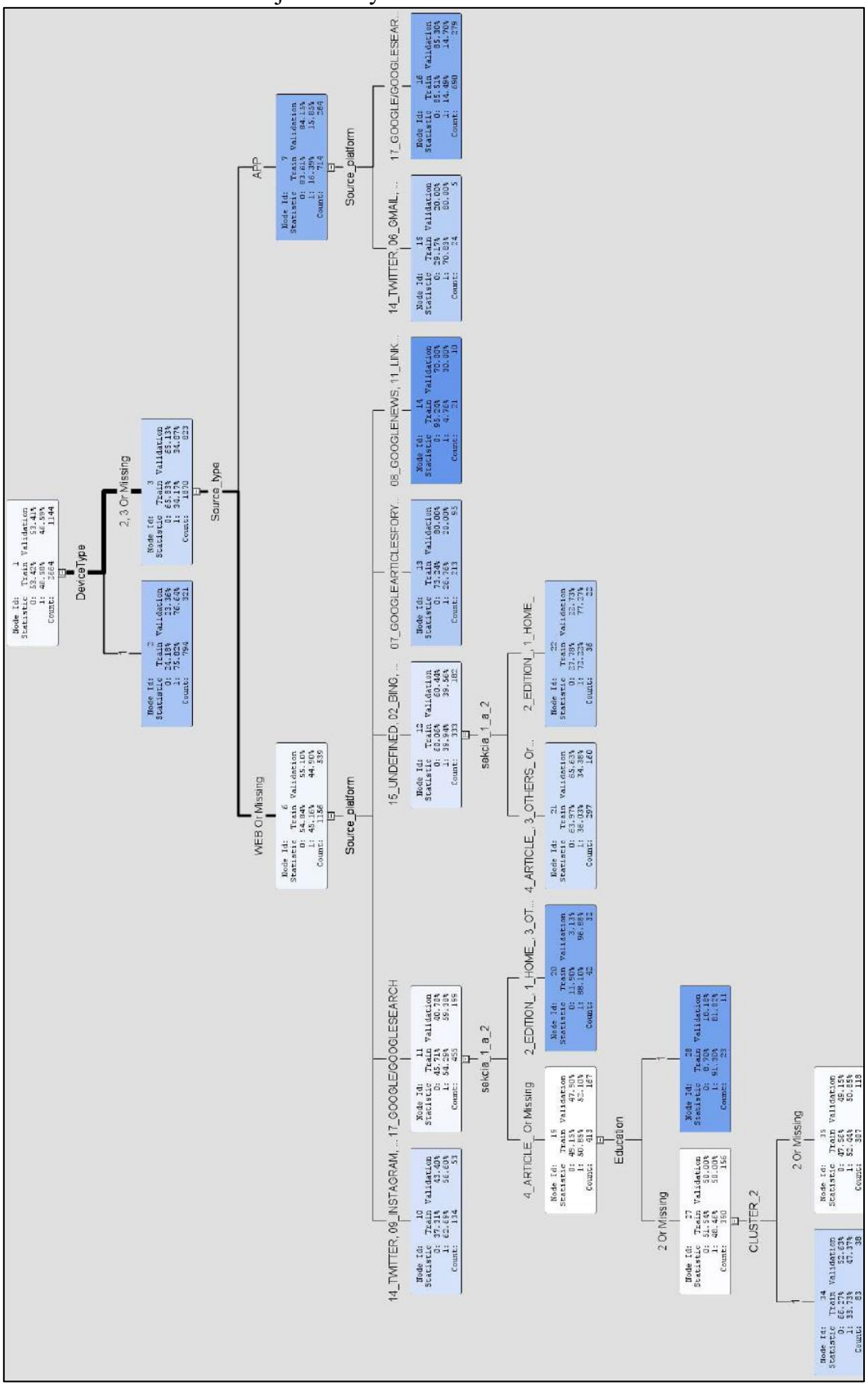
4.4.2 Model rozhodovacieho stromu

Ďalší model, ktorý sme vytvorili, bol model rozhodovacieho stromu. Model rozhodovacieho stromu má dva typy trénovania – autonómne a interaktívne, pričom v našom modeli sme využili práve autonómne trénovanie. Pred vytvorením modelu upravíme nastavenia pre tento model, napr. kritériá na výber vetviacej premennej a nastavenia ovplyvňujúce rast rozhodovacieho stromu.

Vo vlastnostiach modelu je kritérium pre testovacie premenné vybrané podľa toho, o aký typ cieľovej premennej ide, či je intervalová, nominálne (alebo binárna) alebo ordinálna. Keďže naša vysvetľovaná premenná bola binárna, resp. nominálna, pri vytváraní modelu rozhodovacieho stromu sme využili algoritmus CART a testovacie kritérium bola entropia. Niektoré premenné v analyzovanom súbore mali viac ako 2 obmeny, preto sme maximálny počet vetvení na príslušnej úrovni (*Maximum Branch*) nastavili na hodnotu 5. Maximálnu hĺbku stromu (*Maximum Depth*) sme vyskúšali nastaviť na hodnotu 10, avšak rozhodovací strom sa vetvil len do hĺbky 7, preto sme nastavili maximálnu hĺbku na 7. Každá kategoriálna premenná, ktorá mala byť použitá ako testovacie kritérium, musela mať minimálne 2 kategórie (*Minimum Categorical Size*).

Minimálny počet prípadov trénovacej množiny (*Leaf Size*) sme nastavili na hodnotu 20, čo je odporúčaná minimálna veľkosť listového uzla (Mathworks, 2021) a počet rozhodovacích pravidiel (*Number of Rules*) na 20. V tomto nastavení sme sa nechceli obmedzovať, pričom sme nepočítali s tým, že počet rozhodovacích pravidiel pre rozhodovací strom by túto hodnotu prekročil. Po úprave nastavení sme spustili uzol rozhodovacieho stromu.

Obr. č. 12 – Graf stromovej štruktúry



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

Po vytvorení modelu rozhodovacieho stromu dostávame graf stromovej štruktúry (*Tree*, Obr. č. 12), ktorý poskytujú informácie o počte úrovní stromu, počte uzlov (listových aj nelistových), o premenných, ktoré boli použité na jednotlivých úrovniach vetvenia, o početnosti uzlov a pod. Vytvorený model rozhodovacieho stromu obsahuje 6 úrovní a 12 listových a 8 nelistových uzlov.

Farba jednotlivých uzlov súvisí s čistotou uzla. Čím je uzol čistejší, tým je farba tmavšia. Napríklad v uzle číslo 14 pre trénovaciu množinu sa nachádza až 95,24 % prípadov výskytu hodnoty 0 pre danú premennú. Tento uzol je preto najčistejší (Obr. č. 12), čo je vizualizované sýto modrou farbou. V grafe stromovej štruktúry môžeme sledovať aj hrúbku jednotlivých vetiev, ktorá závisí od toho, koľko prípadov bolo z daného uzla odvedených do dcérskeho uzla.

Ako najdôležitejšia premenná pri vytváraní modelu, z hľadiska parametra Importance, bola identifikovaná premenná *Device_Type*, ktorá informuje o type zariadenia použitom na prezeranie článku užívateľom. Najmenej dôležitá je premenná *CLUSTER_2*, ktorá spájala články s tematikou pandémie covid-19. Ďalšie dôležité premenné, ktoré boli využité pri vytváraní modelu rozhodovacieho stromu a teda majú štatisticky významný vplyv na vysvetľovanú premenné, sú premenné *Source_type, Source_platform, Sekcia_1_a_2* a *Education.*

Obr. č. 13 – Uzly a ich čistota (*Tree Leaf Report*)

Tree Leaf Report

Node Id	Depth	Training Observations	Training Percent 1	Validation Observations	Validation Percent 1
2	1	794	0.76	321	0.77
16	3	690	0.14	279	0.15
35	6	307	0.52	118	0.51
21	4	297	0.36	160	0.34
13	3	213	0.27	95	0.20
10	3	134	0.63	53	0.57
34	6	83	0.34	38	0.47
20	4	42	0.88	32	0.97
22	4	36	0.72	22	0.77
15	3	24	0.71	5	0.80
28	5	23	0.91	11	0.82
14	3	21	0.05	10	0.30

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

Informácie o jednotlivých listových uzloch a ich čistote poskytuje *Tree Leaf Report* na Obr. č. 13. Najčistejší listový uzol na trénovacej množine je uzol č. 14, ktorý je tvorený z 95 % prípadmi z triedy 0, čo sme zistili aj na základe Obr. č. 12. Druhý najčistejší uzol pre prípady z triedy 0 je uzol č. 16, ktorý tvorí 86 % prípadov triedy 0. Pre prípady triedy 1 je najčistejší uzol č. 28, ktorý obsahuje 91 % prípadov triedy 1 a druhý najčistejší pre prípady triedy 1 je uzol č. 20, ktorý obsahuje 88 % prípadov triedy 1.

Na základe predikovanej podmienenej pravdepodobnosti sme určili profil užívateľa, ktorý sa s najväčšou pravdepodobnosťou stane predplatiteľom plateného obsahu na spravodajskej webstránke. Takýto užívateľ si prezeral daný článok cez desktop alebo laptop prostredníctvom webovej stránky (nie aplikácie). Na daný článok sa dostal prostredníctvom vyhľadávača Google a tento článok patril do sekcie konkrétnych článkov, nie do podskupín (premenná *Sekcia_1_a_2*). Téma článku bola podľa IAB klasifikácie zaradená do skupiny Vzdelávanie, s čím súhlasí aj tvrdenie pri premennej *CLUSTER_2* – ak užívateľ nepatril do zhluku 2, ktorý spája články s tematikou pandémie covid-19, mal vyššiu pravdepodobnosť predplatiť si platený obsah. Pri pohľade na graf stromovej štruktúry, ide o uzly č. 28, 20 a 2.

Užívateľ, ktorý má najmenšiu pravdepodobnosť predplatiť si platený obsah, podľa modelu rozhodovacieho stromu, si prezeral článok cez webovú stránku prostredníctvom mobilného telefónu. Na článok sa dostal priamo prostredníctvom web stránky thetimes.co.uk alebo prostredníctvom vyhľadávača Google. Čítaný článok nepatril do zhluku č. 2, čiže téma nebola zameraná na pandémiu covid-19 a nepatril ani do kategórie Vzdelávanie (podľa klasifikácie IAB). Pri pohľade na graf stromovej štruktúry, ide o uzly č. 14 a 16.

Vo výstupe sme dostali aj klasifikačnú tabuľku pre vysvetľovanú premennú *Target* pre obe množiny údajov – trénovaciu aj validačnú (Obr. č. 14). Nájdeme tu podiely (vyjadrené v %) správne aj nesprávne klasifikovaných prípadov triedy 0 a 1 pri vstupnej aj predikovanej hodnote premennej *Target*. Na trénovacej množine bolo zo všetkých prípadov triedy 0 71,05 % predikovaných správne a 28,95 % predikovaných nesprávne. Z prípadov triedy 1 bolo 76,40 % predikovaných správne a 23,60 % predikovaných nesprávne. Na validačnej množine bolo z triedy 0 správne predikovaných 73,00 % a z triedy 1 74,50 % prípadov.

Obr. č. 14 – Klasifikačná tabuľka pre vysvetľovanú premennú *Target*

Classification Table

Data Role=TRAIN Target Variable=target Target Label=target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	77.5307	71.0471	1011	37.9505
1	0	22.4693	23.6100	293	10.9985
0	1	30.2941	28.9529	412	15.4655
1	1	69.7059	76.3900	948	35.5856

Data Role=VALIDATE Target Variable=target Target Label=target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	76.6323	72.9951	446	38.9860
1	0	23.3677	25.5159	136	11.8881
0	1	29.3594	27.0049	165	14.4231
1	1	70.6406	74.4841	397	34.7028

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

Ak porovnáme model rozhodovacieho stromu s modelom logistickej regresie, tak na jeho vytvorenie bol použitý menší počet premenných a profil užívateľa sa mierne odlišoval (Tab. č. 6). Podobnosti vidíme v tom, že v oboch prípadoch využil užívateľ na prezeranie článku webovú stránku (nie aplikáciu) cez desktop alebo laptop a téma článku bola podľa IAB klasifikácie zaradená do oblasti Vzdelávania.

Tab. č. 6 – Porovnanie profilu užívateľa pre model logistickej regresie a model rozhodovacieho stromu

Premenná	**Model logistickej regresie**	**Model rozhodovacieho stromu**
Device_Type	Desktop alebo Laptop	Desktop alebo Laptop
Sekcia_1_a_2	Home_	article_
Source_Type	Web	Web
Source_Platform	Gmail	Google/Googlesearch
IAB klasifikácia	Vzdelávanie, Knihy a literatúra, Zdravý životný štýl, Štýl a móda	Vzdelávanie

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

4.4.3 Model náhodného lesa (Model random forest)

Posledným prediktívnym modelom, ktorý sme vytvorili, bol model náhodného lesa. Hlavné tri nastavenia, ktoré sme pred vytvorením modelu upravili, boli *Maximum Number of Trees* – počet rozhodovacích stromov vytvorených v rámci modelu náhodného lesa, ktorý sme nastavili na hodnotu 128. Podľa článku od Oshiro a kol. (2012) má vyšší počet stromov ako 128 štatisticky nevýznamný prínos. *Number of Variables to consider in split search* – počet prediktorov, ktoré sa majú náhodne vybrať v každom uzle a *Proportion of obs in each sample* – počet pozorovaní, ktoré sú vybrané do trénovacej vzorky pri trénovaní každého rozhodovacieho stromu sme nastavili na hodnotu 0,6.

Číselná hodnota pre nastavenie počtu premenných (*Number of Variables to consider in split search*) nebola uvedená, pretože sa líši pre každú množinu údajov. Ladenie a upravovanie druhej mocniny počtu prediktorov riadi mieru korelácie medzi stromami (nižšie hodnoty vedú k menšej korelácii medzi stromami) (Larkin, McManus, 2016).

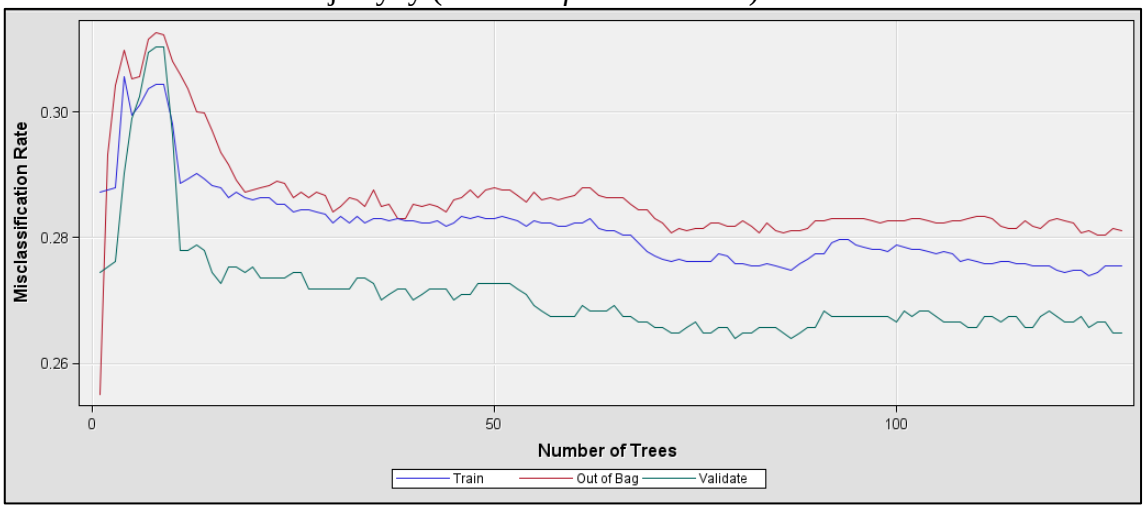
Type of Sample určuje, či sa na určenie trénovacej vzorky použije počet pozorovaní alebo percento pozorovaní. My sme nastavili túto hodnotu na percento pozorovaní (*Proportion*), konkrétne na hodnotu 60 %. Podľa mnohých odporúčaní by mala trénovacia vzorka predstavovať 2/3 pôvodných pozorovaní (Wicklin, 2017).

Ďalším nastavením bolo určenie maximálnej hĺbky uzla v akomkoľvek strome (*Maximum Depth*). Koreňový uzol má hĺbku 0, dcérske uzly koreňového uzla majú hĺbku 1. Hodnotu *Maximum Depth* sme nechceli obmedzovať, takže sme ju najskôr nastavili na 50, aj napriek tomu, že sme nepredpokladali až také vetvenie. Hladinu významnosti (*Significance Level*) sme nastavili na 0,05.

Pri vytváraní modelov rozhodovacích stromov sa využívajú *oob* odhady (pozri časť 3.5) a každý model rozhodovacieho stromu je vytváraný na rovnakom počte pozorovaní ako porovnávacie modely.

Po vytvorení modelu náhodného lesa sme dostali graf celkovej chyby R (*Misclassification rate*, Graf č. 15) a jej hodnoty pri jednotlivých počtoch rozhodovacích stromov. Celková chyba začínala klesať niekde pri 70 stromoch v modeli a najnižšia bola na trénovacej množine pri 125 rozhodovacích stromoch (R = 0,274).

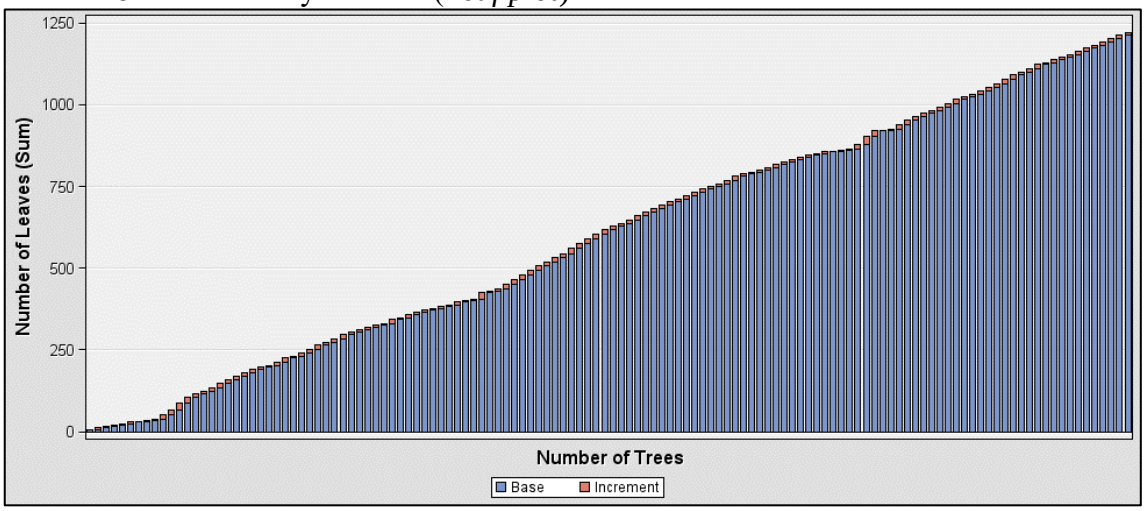
Graf č. 15 – Graf celkovej chyby (*Misclassification rate R*)



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

V grafe listových uzlov (Graf č. 16) vidíme, že prírastok pre každý nový strom je približne rovnaký. To svedčí o tom, že úvodné nastavenia pre vytvorenie modelu náhodného lesa na trénovacej množine boli zvolené správne.

Graf č. 16 – Graf listových uzlov (*Leaf plot*)



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

Výsledky z modelu náhodného lesa vidíme na Obr. č. 15 a Obr. č. 16. Prvá tabuľka obsahuje zhrnutie úvodných nastavení modelu, druhá tabuľka opisuje početnosti trénovacej a validačnej vzorky.

Obr. č. 15 – Informácie o modeli náhodného lesa

Model Information

Parameter	Value	
Variables to Try	7	(Default)
Maximum Trees	128	
Actual Trees	128	
Inbag Fraction	0.6	
Prune Fraction	0	(Default)
Prune Threshold	0.1	(Default)
Leaf Fraction	0.00001	(Default)
Leaf Size Setting	1	(Default)
Leaf Size Used	1	
Category Bins	30	
Interval Bins	100	
Minimum Category Size	5	
Node Size	100000	(Default)
Maximum Depth	50	
Alpha	0.05	
Exhaustive	5000	
Rows of Sequence to Skip	5	(Default)
Split Criterion	.	Gini
Preselection Method	.	Loh
Missing Value Handling	.	Valid value

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

Obr. č. 16 – Počet pozorovaní trénovacej a validačnej množiny

Number of Observations

Type	NTrain	NValid	NTotal
Number of Observations Read	2664	1144	3808
Number of Observations Used	2664	1144	3808

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

Dôležitosť jednotlivých premenných určíme na základe poklesu Giniho indexu (Tab. č. 7). Najvyššia hodnota predstavuje najdôležitejšiu premennú. Medzi najdôležitejších 5 premenných patria premenné: *Source_type, Device_Type, Operating_System, Source_platform* a *sekcia_1_a_2*. Na základe predikovanej podmienenej pravdepodobnosti pre model náhodného lesa, užívateľ, ktorý má najvyššiu pravdepodobnosť predplatiť si platený obsah na spravodajskej webstránke, čítal článok cez desktop alebo laptop v nedeľu alebo utorok v okolo 8 alebo 11 hodiny doobeda. Článok si prezeral cez webstránku (nie cez aplikáciu), konkrétne cez prehliadač Microsoft Edge a článok našiel cez vyhľadávač od spoločnosti Google. Bol zaradený do sekcie Home_ alebo medzi iné (others – napr. sekcia posledných šesť dní). Podľa klasifikácie

IAB možno tento článok zaradiť do kategórie Vzdelávanie a Osobné financie. Rovnako tak tam patria články, ktoré sa spájajú s tematikou pandémie covid-19 (*CLUSTER_2*).

Užívateľ, ktorý má najnižšiu pravdepodobnosť predplatiť si platený obsah, čítal článok cez aplikáciu na mobilnom telefóne alebo tablete, v piatok alebo sobotu o 2 alebo 3 ráno. Na článok sa dostal prostredníctvom aplikácie News od spoločnosti Apple alebo prostredníctvom vyhľadávača Google a prezeral si ho cez prehliadač Chrome. Daný článok patril do sekcie article_, čiže išlo o konkrétny článok. Prezerané články patrili do zhluku č. 3, ktorý spája články o radách a odporúčaniach. Podľa klasifikácie IAB patrili články do kategórie Lekárstvo a zdravie.

Tab. č. 7 – Dôležitosť jednotlivých premenných v modeli náhodného lesa

Variable	Number of Rules	Gini reduction	OOB Gini reduction	Valid Gini reduction
Source_type	142	0.03209	0.03259	0.02978
Device_Type	126	0.02292	0.02292	0.02392
Operating_System	91	0.01082	0.00913	0.01027
Source_platform	119	0.00671	0.00407	0.00495
sekcia_1_a_2	90	0.00664	0.00582	0.01085
sekcia_1	60	0.00418	0.00362	0.00678
Education	70	0.00326	0.00324	0.00312
Affinity_DoW	50	0.00185	0.00089	-0.00005
GeoDMA	35	0.00174	0.00007	0.00079
Affinity_HoD	49	0.00155	-0.00010	-0.00089
Browser	39	0.00120	0.00017	0.00089
Style_and_Fashion	48	0.00093	0.00017	0.00108
Personal_Financial	30	0.00059	0.00033	0.00067
CLUSTER_2	19	0.00039	-0.00014	-0.00035
CLUSTER_5	25	0.00039	-0.00009	0.00001
Books_and_Literature	23	0.00032	-0.00005	-0.00012
Home_and_Garden	18	0.00029	0.00003	-0.00048
Business_and_Financial	14	0.00024	-0.00018	-0.00002
Family_and_Relationships	12	0.00020	-0.00004	0.00023
CLUSTER_3	14	0.00020	-0.00003	0.00006
Medical_Health	10	0.00016	-0.00008	0.00007
CLUSTER_4	9	0.00012	-0.00010	-0.00008
CLUSTER_1	11	0.00012	-0.00011	0.00002
Sports	8	0.00012	-0.00007	-0.00001
Science	10	0.00010	-0.00013	-0.00013
Healthy_Living	10	0.00009	0.00000	0.00000
Events_and_Attractions	8	0.00008	-0.00002	-0.00003
Real_Estate	6	0.00008	-0.00005	-0.00004

Variable	Number of Rules	Gini reduction	OOB Gini reduction	Valid Gini reduction
Music_and_Audio	4	0.00007	-0.00006	-0.00003
Travel	6	0.00006	-0.00007	-0.00002
Hobbies_and_Interests	5	0.00005	-0.00003	-0.00006
Careers	2	0.00003	0.00000	-0.00001
Automotive	2	0.00003	-0.00001	0.00000
Technology_and_Computing	3	0.00002	-0.00004	-0.00003
Television	3	0.00002	0.00001	-0.00004
Shopping	1	0.00001	-0.00001	-0.00001
Fine_Art	1	0.00001	-0.00004	-0.00002
Pets	1	0.00001	-0.00002	0.00000
Pop_Culture	1	0.00001	-0.00001	0.00000
Food_and_Drink	1	0.00001	-0.00003	-0.00004

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

Rovnako ako v prípade predchádzajúceho modelu, aj tu sme dostali klasifikačnú tabuľku pre vysvetľovanú premennú *Target* (Obr. č. 17).

Obr. č. 17 – Klasifikačná tabuľka pre vysvetľovanú premennú *Target*

Classification Table

Data Role=TRAIN Target Variable=target Target Label=target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	69.9479	84.8911	1208	45.3453
1	0	30.0521	41.8211	519	19.4820
0	1	22.9456	15.1089	215	8.0706
1	1	77.0544	58.1789	722	27.1021

Data Role=VALIDATE Target Variable=target Target Label=target

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	70.6434	86.2520	527	46.0664
1	0	29.3566	41.0882	219	19.1434
0	1	21.1055	13.7480	84	7.3427
1	1	78.8945	58.9118	314	27.4476

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

Tabuľka zobrazuje početnosti a percentuálne početnosti pre pozorované a predikované hodnoty vysvetľovanej premennej. V porovnaní s predchádzajúcim

modelom rozhodovacieho stromu (DT) je v modeli náhodného lesa (RF) presnejšia predikcia prípadov triedy 0 (trénovacia množina DT = 71,05 % vs. RF = 84,89 %, validačná množina DT = 73,00 % vs. 86,25 %) a nepresnejšia predikcia prípadov triedy 1 (trénovacia množina DT = 76,39 % vs. RF = 58,18 %, validačná množina DT = 74,48 % vs. RF = 58,91 %).

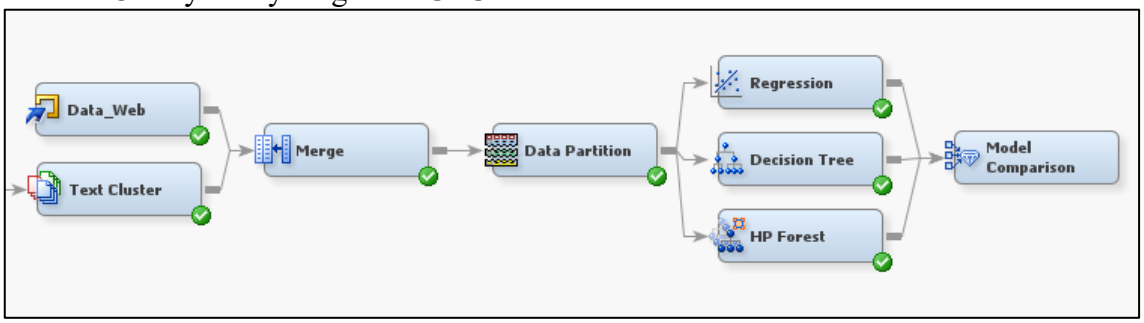
Na základe získaných výsledkov z troch prediktívnych modelov sme porovnali tieto modely, vyhodnotili ich kvalitu a vybrali ten, ktorý najlepšie predikuje hodnoty vysvetľovanej premennej.

4.5 Validácia prediktívnych modelov

V predchádzajúcej podkapitole sme odhadli tri prediktívne modely – model binárnej logistickej regresie, model rozhodovacieho stromu a model náhodného lesa. Modely boli vytvárané na trénovacej množine, ktorá tvorila 70 % vstupných údajov. Kvalitu týchto modelov sme otestovali na validačnej množine, ktorá tvorí zvyšných 30 % zo vstupných údajov, ktoré neboli použité na vytvorenie prediktívnych modelov. Cieľom bolo zistiť, či niektorý z modelov nebol preučený – dosahoval oveľa lepšie výsledky na trénovacej množine ako na validačnej.

SAS EM ponúka uzol Model Comparison, ktorý poslúžil na výpočet charakteristík porovnávajúcich kvalitu modelov, na porovnanie jednotlivých modelov a na následný výber modelu s najlepšou prediktívnou schopnosťou. Výsledný diagram v SAS EM vidíme na Obr. č. 18. Ako kritérium na výber najlepšieho modelu sme zvolili ROC index.

Obr. č. 18 – Výsledný diagram v SAS EM

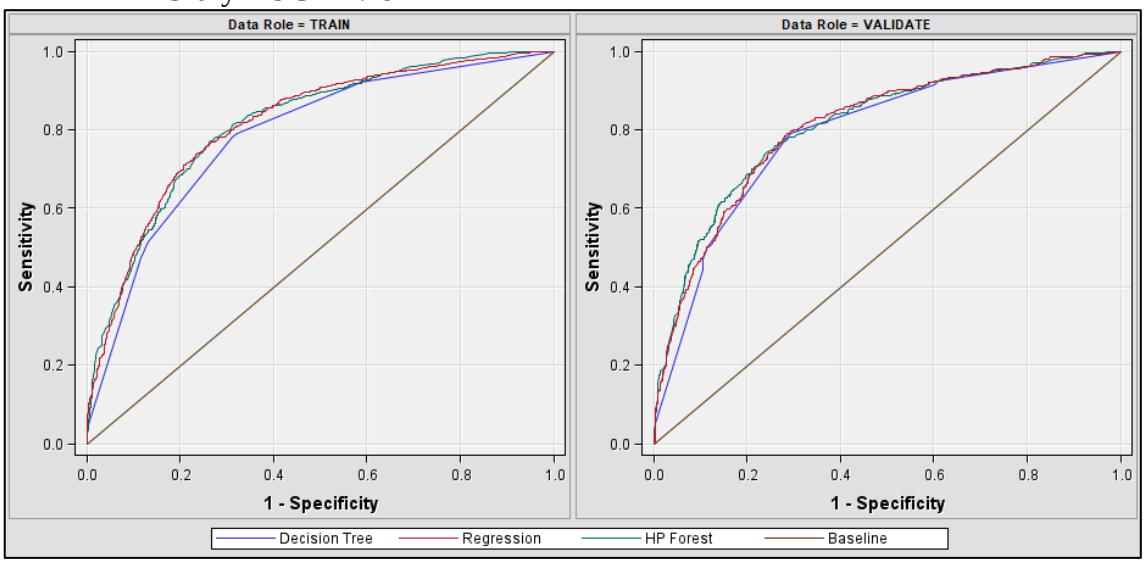


Zdroj: GroupM data source – thetimes.co.uk website; Vlastné spracovanie v SAS EM

Vo výsledkoch dostávame dva grafy ROC kriviek pre vysvetľovanú premennú – jeden z trénovacej a jeden z validačnej množiny (Graf č. 17).

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. Pre jednotlivé modely bola hranica klasifikácie vysvetľovanej premennej nastavená na hodnotu 0,5.

Graf č. 17 – Grafy ROC kriviek



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

Tab. č. 8 – Hodnotenie modelov na trénovacej a validačnej množine

Model	**ROC index**	**Celková chyba *R***	**MSE**	**Celková správnosť *A***	**Presnosť**	**Senzitívnosť**	**1 – Špecifičnosť**	***F*-skóre**
REG_TRAIN	0.815	0.2492	0.1741	0.7508	0.7394	0.7180	0.2207	0.7285
RF_TRAIN	0.818	0.2755	0.1854	0.7245	0.7705	0.5818	0.1511	0.6630
DT_TRAIN	0.789	0.2646	0.1828	0.7354	0.6971	0.7639	0.2895	0.7290
REG_VALID	0.807	0.2587	0.1780	0.7413	0.7215	0.7242	0.2439	0.7228
RF_VALID	0.808	0.2649	0.1870	0.7351	0.7889	0.5891	0.1375	0.6745
DT_VALID	0.792	0.2631	0.1815	0.7369	0.7064	0.7448	0.2700	0.7251

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

V Tab. č. 8 sú zhrnuté všetky vypočítané charakteristiky pre vytvorené prediktívne modely – model binárnej logistickej regresie (REG), model rozhodovacieho stromu (DT) a model náhodného lesa (RF) na trénovacej (*_TRAIN) aj validačnej (*_VALID) množine. Okrem základných štatistík ako celková chyba a priemerná štvorcová chyba MSE (*Mean squared error*) tabuľka obsahuje aj celkovú správnosť, presnosť, senzitívnosť, 1 – špecifičnosť a *F*-skóre.

Podľa ROC indexu ako najlepší prediktívny model vychádza na validačnej množine model náhodného lesa, hoci aj ostatné modely majú podľa ROC indexu dobrú predikčnú schopnosť, čo dokazujú aj ostatné hodnotiace kritériá. Napríklad podľa *F*-skóre a celkovej správnosti vychádza najlepšie model rozhodovacieho stromu.

Ak berieme ako rozhodovacie kritérium ROC index, tak najlepší prediktívny model, podľa trénovacej množiny, je model náhodného lesa. Tento model najlepšie predikuje predplatiteľov plateného obsahu spravodajskej webstránky. Veľkosť ROC indexu na trénovacej množine bola 81,80 % a na validačnej 80,80 %. Celková chyba *R* na trénovacej množine bola 27,55 % a na validačnej sa znížila na hodnotu 26,49 %. *F*-skóre sa na validačnej množine mierne zvýšilo z hodnoty 66,30 % na 67,45 %, teda o 1,15 percentuálneho bodu.

Jednotlivé modely viedli k podobným výsledkom, najviac premenných vstupovalo do modelu náhodného lesa. Vo všetkých troch modeloch patrili medzi najdôležitejšie premenné *Source_type, Device_Type, Source_Platform, Sekcia_1_a_2, Browser* a *Education*, čo znamená, že pri predplatení si plateného článku na spravodajskej webstránke najviac zavážili fakty, či prišiel užívateľ na danú webstránku z webovej stránky alebo z aplikácie (*Source_type*), aký typ zariadenia využíval – desktop alebo laptop, mobilný telefón alebo tablet (*Device_type*), aká bola referenčná stránka, z ktorej užívateľ prišiel (*Source_platform*), do akej sekcie bol článok zaradený (*Sekcia_1_a_2*) a aký webový prehliadač užívateľ využíval (*Browser*). Po obsahovej stránke zaujali užívateľov webstránky, ktoré patrili do kategórie Vzdelávania (*Education*). Pri modeloch logistickej regresie a náhodného lesa záviselo aj na tom, v ktorý deň a v akom čase si užívateľ článok prezeral.

Profily užívateľov, ktorí sa s najväčšou pravdepodobnosťou stanú predplatiteľmi zamknutého obsahu, podľa jednotlivých modelov sa nachádzajú v Tab. č. 9 a profily užívateľov s najnižšou pravdepodobnosťou predplatenia si plateného obsahu sú v Tab. č. 10.

Ako vidíme v tabuľkách nižšie, profily užívateľov, získané z jednotlivých modelov, sa od seba veľmi nelíšia, čo však nevadí, pretože cieľom bolo identifikovať profil užívateľa patriaceho do cieľovej skupiny. To, že sme si pomocou rôznych modelov potvrdili podobné výsledky dokazuje, že identifikovanie profilov je správne.

Tab. č. 9 – Profily užívateľov s najvyššou pravdepodobnosťou predplatenia si plateného obsahu

Premenná	Model logistickej regresie	Model rozhodovacieho stromu	Model náhodného lesa
Source_Type	Web	Web	Web
Device_Type	Desktop alebo Laptop	Desktop alebo Laptop	Desktop alebo Laptop
Source_Platform	Gmail	Google/Googlesearch	Google/Googlesearch
Sekcia_1_a_2	Home_	article_	Home_, others_
Browser	Edge	X	Edge
IAB klasifikácia	Vzdelávanie, Knihy a literatúra, Zdravý životný štýl, Štýl a móda	Vzdelávanie	Vzdelávanie, Osobné financie
Hour of Day	11	X	8, 11
Day of Week	Utorok	X	Nedeľa, Utorok

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

Tab. č. 10 – Profily užívateľov s najnižšou pravdepodobnosťou predplatenia si plateného obsahu

Premenná	Model logistickej regresie	Model rozhodovacieho stromu	Model náhodného lesa
Source_Type	Aplikácia	Web	Aplikácia
Device_Type	Tablet	Mobilný telefón	Mobilný telefón alebo Tablet
Source_Platform	Google News	internal	Google, Apple News
Sekcia_1_a_2	article_	edition_, others_	article_
Browser	Samsung Browser	X	Chrome
IAB klasifikácia	Lekárstvo/ Zdravie	X	Lekárstvo/ Zdravie
Hour of Day	3	X	2, 3
Day of Week	Piatok	X	Piatok, Sobota

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

Užívateľ, ktorí sa s najväčšou pravdepodobnosťou stane predplatiteľom zamknutého obsahu, si prezeral článok v utorok alebo v nedeľu o 8 alebo 11 hodine. Článok si prezeral cez desktop alebo laptop, cez webovú stránku a prehliadač od spoločnosti Microsoft - Edge a prišiel naň cez vyhľadávať od Google alebo cez ich mailového klienta – Gmail. Článok patril buď do sekcie konkrétnych článkov (article_) alebo do sekcie Home_ alebo others_. Bol zameraný na Vzdelávanie, prípadne na Zdravý životný štýl, Osobné financie, Knihy a literatúra alebo Štýl a móda.

Užívateľ, ktorý sa s najnižšou pravdepodobnosťou stane predplatiteľom plateného obsahu, si prezeral článok v piatok alebo v sobotu, o 2 alebo 3 hodine ráno. Tento článok si prezeral cez mobilnú aplikáciu prostredníctvom mobilného telefónu alebo tabletu. Článok patril medzi konkrétne články – sekcia article_, prípadne edition_ alebo others_. Na článok sa dostal prostredníctvom Google News, Apple News, cez vyhľadávač od Google alebo priamo cez spravodajskú webstránku. Téma článku bola zameraná na Lekárstvo alebo Zdravie.

Všetky tri prediktívne modely poskytli podobné výsledky a predikovali užívateľov dobre. Ak sa rozhodneme využiť ten model, ktorý získal najlepšie hodnoty podľa mier slúžiacich na hodnotenie modelu – a to model náhodného lesa, tak tento výsledný model je následne možné aplikovať na cielenie v online reklamnej kampani. Podľa získaných charakteristík by mal takýto model najlepšie predikovať užívateľov, ktorí sa stanú predplatiteľmi plateného obsahu na spravodajskej webstránke. Reklamnou kampaňou chceme osloviť takých užívateľov, ktorí majú podobné alebo rovnaké behaviorálne znaky, ako návštevníci webstránky, ktorí už prejavili o články záujem a následne sa stali predplatiteľmi. Takýchto užívateľov potrebujeme osloviť online reklamnou kampaňou kdekoľvek na internete sa nachádzajú, a to práve prostredníctvom programatického nákupu reklamy. Keď užívateľ príde na akúkoľvek stránku, na ktorej je reklamný priestor a spĺňa kritériá, ktoré vyhodnotil model ako dôležité pre užívateľov s najvyššou pravdepodobnosťou predplatenia si plateného obsahu, zobrazí sa danému užívateľovi reklama, ktorá bude súčasťou reklamnej kampane zameranej na získanie nových predplatiteľov. Takéto cielenie reklamy na užívateľov je nákladovo efektívne, pretože sa reklama nezobrazuje užívateľom, pre ktorých by bola reklama irelevantná alebo nežiaduca. Tým sa zvyšuje efektivita cielenia reklamnej kampane a zároveň sa znižujú náklady na reklamu a zvyšuje zisk.

5 Diskusia

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. Vysoký dôraz sa kladie najmä na personalizáciu jednotlivých ponúk, ktoré sú bežným spotrebiteľom každý deň podsúvané. Jedným zo spôsobov ako zaujať spotrebiteľov a odprezentovať svoj produkt je reklama. 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.

Na tému cielenia online reklamy a využívania digitálneho a online marketingu bolo realizovaných niekoľko výskumov. De Bock a Van den Poel (2010) sa vo svojej práci zamerali na využitie demografického cielenia a podobne ako v našej práci, aj oni sa snažili predikovať profily návštevníkov internetovej stránky, ktoré možno použiť na účely cielenia online reklamy. V uvedenej práci bol model náhodného lesa porovnávaný s algoritmami CART, C4.5, Bagging a AdaBoost. Z uvedených algoritmov vyšiel najlepšie model náhodného lesa. Získané výsledky a publiká bolo následne možné využiť pri výbere webových stránok, ktoré sa majú použiť na online reklamu. Výskum, ktorý realizovali Bogaert a kol. (2017), bol zameraný na cielenie online reklamy prostredníctvom sociálnej siete Facebook na užívateľov, ktorí by si mohli kúpiť produkty súvisiace s futbalom. Na predikciu bolo využitých až šesť algoritmov – model logistickej regresie, model náhodného lesa, Adaboost, kernel factory, neurónovú sieť a model rotation forest. . Z hľadiska presnosti (Accuracy) boli najlepšie modely Adaboost a model logistickej regresie.

Aj mnohé ďalšie výskumy poukazujú na to, že oblasť strojového učenia a marketingu sa v posledných rokoch začínajú úzko prepájať. Brei sa vo svojom výskume (Brei, 2020) zameral na opis metód strojového učenia s prepojením na oblasť marketingu a aký význam má strojové učenie pre marketing a spotrebiteľov, pričom sa snaží poukázať na zovšeobecniteľnosť metód strojového učenia, na základe vedeckých objavov a ich aplikovateľnosť aj v iných odvetviach, vrátane marketingu. Publikácia tiež obsahuje

analýzu aplikácií strojového učenia publikovaných v marketingových a manažérskych časopisoch, knihách a iných štúdiách. Kniha od Syam a Kaul (2021) je tvorená názormi dvoch odborníkov, ktorí sa roky zaoberajú strojovým učením – jeden z teoretického hľadiska a druhý z hľadiska aplikácie modelov strojového učenia v oblasti marketingu a predaja. Ich cieľom bolo prepojiť koncepty, ktoré tvoria základ metód strojového učenia s praktickou aplikáciou v oblasti marketingu a preklenúť tak priepasť medzi dátovými vedcami alebo akademickými odborníkmi a business odborníkmi z praxe. Keďže si uvedomovali komplexnosť danej témy, rozhodli sa zamerať na tri metódy strojového učenia využívané v oblasti marketingu a predaja, a to neurónové siete, metódu podporných vektorov (*SVM*) a náhodný les (*random forest*). Poukazujú aj na to, že dané modely boli overené z hľadiska ich použiteľnosti pri riešení širokej škály marketingových problémov. K využitiu strojového učenia v oblasti marketingu vyzývali vo svojom výskume aj Ma a Sun (2020), napríklad pri mapovaní nákupných ciest zákazníkov a pri podpore rozhodovania v súvislosti s marketingovou stratégiou. Náš výskum využíval behaviorálne cielenie a podľa Chena a Stallaerta (2014) sa online reklamná kampaň, ktorá je založená na behaviorálnom cielení, stáva významnou časťou online marketingu. V ich príspevku sú analyzované aj ekonomické dôsledky tohto cielenia na hlavných aktérov (zadávateľov reklamy a poskytovateľov reklamného priestoru). De Bock a Van den Poel (2010) sa zasa zamerali na demografické cielenie a predikciu profilov používateľov internetovej stránky pomocou metódy náhodného lesa (*random forest*).

Ako opisujú aj jednotlivé štúdie a výskumy, v súčasnosti sa začínajú metódy strojového učenia aplikovať aj v oblasti marketingu, resp. v online marketingu. V našej práci sme sa rozhodli zamerať na tri metódy, ktoré dosahujú dobré výsledky, čo potvrdzuje aj literatúra. Podobne ako pri spomenutých výskumoch, aj naším cieľom bolo identifikovanie užívateľov, ktorí patria do cieľovej skupiny prostredníctvom vytvorenia prediktívnych modelov za použitia metód strojového učenia. Cieľová skupina je charakterizovaná určitými behaviorálnymi znakmi a má na ňu byť následne cielená online reklama. Spravodajská webstránka ponúka články na čítanie. Obsah týchto článkov však nie je zadarmo. Na webstránke je nutná registrácia a následná možnosť skúšobnej verzie, kedy sú články sprístupnené na mesiac zadarmo (tzv. trial verzia). Po uplynutí tejto doby musí užívateľ za prístup k článkom zaplatiť poplatok v podobe ročného predplatného alebo sa mu prístup k článkom skončí. Cieľom spravodajskej webstránky je osloviť reklamnou kampaňou takých užívateľov, ktorí sa s najväčšou pravdepodobnosťou stanú predplatiteľmi plateného obsahu na tejto webstránke (po tom, ako im uplynie skúšobná

verzia). Zdrojové údaje pochádzali z marketingovej spoločnosti GroupM Slovakia a obsahovali informácie o 3808 užívateľov spravodajskej webstránky thetimes.co.uk.

V práci sme zosumarizovali tri metódy strojového učenia, konkrétne model binárnej logistickej regresie, model rozhodovacieho stromu a model náhodného lesa a tieto metódy boli následne aplikované pri vytvorení prediktívnych modelov a predikovaní užívateľov spravodajskej webstránky, ktorí sa s najväčšou pravdepodobnosťou stanú predplatiteľmi zamknutého obsahu.

Naplnenie hlavného cieľa dizertačnej práce bolo realizované prostredníctvom viacerých krokov. Prvým z nich spracovanie aktuálnych informácií a poznatkov o digitálne marketingu a reklamnej online kampane. Tá sa rozdeľuje na niekoľko typov, pričom ide o typy z hľadiska cielenia online reklamy a z hľadiska nákupu online reklamy. V našom prípade sme sa primárne zameriavali na programatický nákup reklamy, najmä z dôvodu využitia vytvorených prediktívnych modelov pri takomto nákupe online reklamy a pri následnom cielení.

Ďalším krokom bolo oboznámenie sa s metodikou vytváraných prediktívnych modelov a s oblasťou strojového učenia. Súčasťou metodiky bolo aj oboznámenie sa s rôznymi mierami, ktoré sa využívajú na hodnotenie modelov. Pomocou týchto mier bol vybraný model, ktorý najlepšie predikoval cieľovú skupinu. Medzi tieto miery patrili klasifikačné miery, ROC krivka a *F*-skóre.

Pri praktickej aplikácii bolo najskôr nutné upraviť dáta tak, aby sa mohli stať vstupnými údajmi do prediktívnych modelov. Príprava dát spočívala v úprave existujúcich premenných na nové premenné a súčasťou prípravy bola aj analýza textu, tzv. Text Miningu, kedy sme webovú adresu čítaného článku rozanalyzovali a vytvorili tak päť zhlukov, ktoré boli použité ako vstupné premenné pri prediktívnych modeloch. Vytváranie prediktívnych modelov pozostávalo z vhodne zvolených nastavení parametrov pre jednotlivé modely a ich detailného opisu. Pomocou softvéru SAS sme vytvorili tri prediktívne modely – model binárnej logistickej regresie, model rozhodovacieho stromu a model náhodného lesa. Práca je teda zároveň ukážkou aplikácie zvolených metód strojového učenia v softvéri SAS. Vytvorené modely boli porovnané prostredníctvom viacerých mier využívaných na hodnotenie modelov.

Získané výsledky identifikovali faktory, ktoré signifikantne ovplyvňujú pravdepodobnosť, že si užívateľ predplatí alebo nepredplatí platený obsah na spravodajskej webstránke. Rovnako tak sme získali profily užívateľov, ktorí sú najviac náchylní na to, predplatiť si zamknutý obsah.

Dizertačná práca poskytuje návod ako využiť dáta o užívateľoch rôznych webstránok na vytvorenie prediktívnych modelov, ktoré budú predikovať užívateľov alebo návštevníkov, ktorí sa s najväčšou pravdepodobnosťou stanú predplatiteľmi alebo na danej webstránke nakúpia. Aj keď vytvorený prediktívny model nie je univerzálnym modelom pre každú webovú stránku, postupnosť krokov využitých, pri jeho tvorbe, je možné aplikovať aj pre iné webové stránky a následné reklamné kampane.

Hlavné výsledky práce, ktoré sme získali prostredníctvom jednotlivých analýz, sú zhrnuté v časti Záver.

Záver

Spoločnosti často disponujú veľkým množstvom údajov, z ktorých však nedokážu prinášať užitočné závery, pokiaľ dáta nie sú vhodne spracované. Informácie o návštevnosti webovej stránky poskytujú majiteľom webových stránok príležitosť na identifikáciu tých užívateľov, ktorí sú pre nich najviac ziskoví a na základe ich správania nájsť užívateľov, ktorí by sa potenciálne mohli stať platenými zákazníkmi.

Cieľom dizertačnej práce bolo identifikovanie profilu internetových užívateľov, ktorí patria do cieľovej skupiny, ktorá je charakterizovaná určitými behaviorálnymi znakmi a na ktorú má byť cielená online reklama, pomocou prediktívneho modelovania.

V dizertačnej práci sme sa zamerali na analýzu informácií o návštevníkoch spravodajskej webstránky thetimes.co.uk, ktorí na danej stránke čítali rôzne články a na predikciu tých, ktorí by sa s najväčšou pravdepodobnosťou mohli stať predplatiteľmi zamknutého obsahu na tejto stránke. Vytvoreniu prediktívnych modelov predchádzala úprava vstupných údajov, ktorá okrem vytvárania nových premenných zahŕňala aj analýzu textu – Text Mining. Prediktívne modely, ktoré boli v práci vytvorené, sú model binárnej logistickej regresie, model rozhodovacieho stromu a model náhodného lesa. Všetky tri modely boli následne porovnávané prostredníctvom mier klasifikácie, ROC indexu a *F*-skóre.

Na základe získaných výsledkov z prediktívnych modelov môžeme konštatovať, že všetky tri prediktívne modely predikovali cieľovú skupinu veľmi dobre a odlišnosti v metrikách slúžiacich na ich porovnanie boli len minimálne. Podľa ROC indexu bol najlepší prediktívny model na validačnej množine model náhodného lesa, kde bola hodnota; ROC indexu 0,808. Avšak ani výsledky z ostatných modelov nevyšli oveľa horšie – pre model logistickej regresie to bola hodnota 0,807 a pre model rozhodovacieho stromu 0,792.

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*), 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*). Pri modeloch logistickej regresie a náhodného lesa záviselo aj na tom, v ktorý deň a o akom čase si užívateľ článok prezeral.

Podľa uvedených faktorov užívateľ, ktorý sa s najväčšou pravdepodobnosťou stane predplatiteľom plateného obsahu na spravodajskej webstránke, si daný článok prezeral v utorok alebo v nedeľu, okolo 8 alebo 11 hodiny ráno. Pri čítaní využil webstránku namiesto aplikácie, desktop alebo laptop a prehliadač od spoločnosti Microsoft – Edge. Na stránku prišiel prostredníctvom vyhľadávača od spoločnosti Google alebo cez ich mailového klienta Gmail. Článok bol zahrnutý do sekcie article_, Home_ alebo others_ a témou bolo Vzdelávanie, Osobné financie, Knihy a literatúra, Zdravý životný štýl a Štýl a móda.

Užívateľ, ktorý by si predplatil platený obsah s najnižšou pravdepodobnosťou, čítal článok v piatok alebo v sobotu, o 2 alebo 3 hodine ráno. Využil na to mobilný telefón alebo tablet, aplikáciu a prehliadač Samsung Browser alebo Chrome. Na článok sa dostal prostredníctvom aplikácie News od spoločnosti Google alebo Apple, cez vyhľadávač od spoločnosti Google alebo priamo cez stránku thetimes. Článok bol zaradený do sekcie article_, edition alebo others_ a témou bolo Lekárstvo a Zdravie.

Získané modely je možné využiť na cielenie v online reklamnej kampani. Podľa zistených charakteristík by mal prediktívny model najlepšie predikovať užívateľov, ktorí sa s najvyššou pravdepodobnosťou stanú predplatiteľmi plateného obsahu na spravodajskej webstránke. Reklamná kampaň bude teda zameraná na takých užívateľov, ktorí majú podobné alebo rovnaké behaviorálne charakteristiky, ako návštevníci webstránky, ktorí prejavili záujem o články a ktorí sa následne stali predplatiteľmi. Keďže potrebujeme takýchto užívateľov osloviť online reklamnou kampaňou kdekoľvek na internete sa nachádzajú, na nákup reklamy bude využitý programatického nákupu, ktorý zobrazí reklamu takým návštevníkom, ktorých model vyhodnotil ako relevantných pre zhliadnutie danej reklamy, a to užívateľov s najvyššou pravdepodobnosťou predplatenia si plateného obsahu. Využitie modelov strojového učenia v spojitosti s online marketingom prináša nielen efektivitu z hľadiska cielenia reklamy a personalizácie reklamného obsahu, ale aj z hľadiska šetrenia nákladov a zvýšenia zisku.

Zoznam použitej literatúry

ADIL, M., 2022. *10+ advantages and disadvantages of online digital marketing* [online]. [cit. 30. apríla 2022]. Dostupné na internete: https://adilblogger.com/advantages-disadvantages-online-digital-marketing/.

ALLISON, P. D., 2012. *Logistic Regression Using SAS®: Theory and Application, Second Edition*. Cary, NC: SAS Institute Inc., 348 s. ISBN 978-1-59994-641-2.

ANIRUDH, V. K., 2019. *What Is Machine Learning: Definition, Types, Applications and Examples* [online]. [cit. 4. decembra 2020]. Dostupné na internete: https://www.toolbox.com/tech/artificial-intelligence/tech-101/what-is-machine-learning-definition-types-applications-and-examples/#_003.

ARIF, R., 2020. *A Simple Introduction to The Random Forest Method* [online]. [cit. 10. januára 2021]. Dostupné na internete: https://arifromadhan19.medium.com/a-simple-introduction-to-the-random-forest-method-badc8ee6c408.

BALODI, T., 2020. *What is Stemming and Lemmatization in NLP?* [online]. [cit. 22. októbra 2021]. Dostupné na internete: https://www.analyticssteps.com/blogs/what-stemming-and-lemmatization-nlp.

BERK, R. A., 2016. *Statistical Learning from a Regression Perspective, Second Edition*. Springer Texts in Statistics, 372 s. ISBN 978-3-319-44047-7.

BERKA, P., 2003. *Dobývání znalostí z databází*. Praha: Academia, 366 s. ISBN 80-200-1062-9.

BERRY, M. J. A., LINOFF, G. S, 2004. *Data mining Techniques*. *For Marketing, Sales, and Customer Relationship management.* Second Edition. Indiana: Wiley Publishing, Inc., Indianapolis, 643 s. ISBN 978-0-471-47064-9.

BOGAERT, M. et al., 2017. Identifying soccer players on Facebook through predictive analytics. *Decision Analysis,* 14(4), 274-297.

BOSE, I., CHEN, X., 2009. Quantitative model for direct marketing: A review from systems perspective. *European Journal of Operational Research,* 195(1), 1-16.

BREI, V., 2020. Machine Learning in Marketing: Overview, Learning Strategies, Applications, and Future Developments. *Foundations and Trends® in Marketing*, 14(3), 173-236.

BREIMAN, L., 2001. Random Forests. Machine Learning. 45(1). 5-32. 10.1023/A:1010950718922.

BROWNLEE, J., 2016. *Supervised and Unsupervised Machine Learning Algorithms* [online]. [cit. 9. novembra 2020]. Dostupné na internete: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

BROWNLEE, J., 2019. *14 Different Types of Learning in Machine Learning* [online]. [cit. 16. novembra 2020]. Dostupné na internete: https://machinelearningmastery.com/types-of-learning-in-machine-learning/.

CHAKRABORTY, G., PAGOLU, M., GARLA, S., 2013. Text mining and analysis: Practical methods, examples, and case studies using Sas. Cary, NC: SAS Institute Inc., 340 s. ISBN 978-1-61290-551-8.

CHEN, J., STALLAERT, J., 2014. An economic analysis of online advertising using behavioral targeting. *MIS Quarterly*, 38(2), 429-449.

CONSTANTIN, C., 2015. Using the Logistic Regression model in supporting decisions of establishing marketing strategies. *Economic Sciences*, 8(57).

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

DALHOUSIE UNIVERSITY, 2012. *SEC. 16.4 convergence - Dalhousie University* [online]. [cit. 23. októbra 2021]. Dostupné na internete: http://web.cs.dal.ca/~anwar/ir/lecturenotes/l7.pdf.

DATA-FLAIR, 2020. *Machine Learning Tutorial – All the Essential Concepts in Single Tutorial* [online]. [cit. 7. decembra 2020]. Dostupné na internete: https://data-flair.training/blogs/machine-learning-tutorial/.

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.

FAKIR, Y., AZALMAD, M., ELAYCHI, R., 2020. Study of The ID3 and C4.5 Learning Algorithms. *Journal of Medical Informatics and Decision Making,* 1(2), 29-43.

GAO, J., BUFFALO, S., 2012. *Clustering Lecture 2: Partitional Methods* [online]. [cit. 23. októbra 2021]. Dostupné na internete: https://cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_partitional.pdf.

HERNÁNDEZ, I. et al., 2012. A statistical approach to URL-based web page clustering. *WWW '12 Companion: Proceedings of the 21st International Conference on World Wide Web*, 525-526.

IAB SLOVAKIA, 2020. *Výdavky do internetovej reklamy vzrástli v roku 2019 o 7%* [online]. [cit. 2. februára 2021]. Dostupné na internete: https://www.iabslovakia.sk/tlacove-centrum/vydavky-do-internetovej-reklamy-vzrastli-v-roku-2019-o-7/.

IAB SLOVAKIA, 2021. *Reporty návštevnosti a sociodemografie* [online]. [cit. 10. februára 2021]. Dostupné na internete: https://www.iabslovakia.sk/iabmonitor/reporty-navstevnosti-sociodemografie/.

IAB SLOVAKIA, 2022. *Návštevnosť Slovenského Internetu v Roku 2021* [online]. [cit. 28. apríla 2022]. Dostupné na internete: https://www.iabslovakia.sk/iabmonitor/navstevnost-slovenskeho-internetu-v-roku-2021/.

IBM CLOUD EDUCATION, 2020. *What is text mining?* [online]. [cit. 12. októbra 2021]. Dostupné na internete: https://www.ibm.com/cloud/learn/text-mining.

IGI-GLOBAL, 2022. *What is internet penetration* [online]. [cit. 28. apríla 2022]. Dostupné na internete: https://www.igi-global.com/dictionary/internet-penetration/15438.

INFOLEARNERS, 2022. *Online marketing advantages and disadvantages* [online]. [cit. 30. apríla 2022]. Dostupné na internete: https://infolearners.com/online-marketing-advantages-and-disadvantages/

KARIM, M., RAHMAN, R. M., 2013. Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. *Journal of Software Engineering and Applications,* 6(4), 196-206.

KOMPRDOVÁ, K., 2012a. *Pokročilé neparametrické metody* [online]. [cit. 11. januára 2021]. Dostupné na internete: https://is.muni.cz/el/1431/jaro2012/Bi7490/um/Rozhodovaci_lesy2012_tisk.pdf.

KOMPRDOVÁ, K., 2012b. *Rozhodovací stromy a lesy.* Vol. 1. Brno: Akademické nakladatelství CERM, 99 s. ISBN 978-80-7204-785-7.

KORSTANJE, J., 2021. *The F1 score* [online]. [cit. 14. septembra 2021]. Dostupné na internete: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.

KUMAR, N., 2019. *Advantages and Disadvantages of Random Forest Algorithm in Machine Learning* [online]. [cit. 11. januára 2021]. Dostupné na internete: http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html.

LARKIN, T. K., MCMANUS, D. J., 2016. *Social Media, Anonymity, and Fraud: Forest Node in SAS Enterprise Miner* [online]. [cit. 26. októbra 2021]. Dostupné na internete: https://www.lexjansen.com/wuss/2016/60_Final_Paper_PDF.pdf.

LI, B. et al., 2019. Incorporating URL embedding into ensemble clustering to detect web anomalies. *Future Generation Computer Systems*, vol 96, 176-184.

LIN, W., et al., 2017. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. *IEEE Access,* 5, 16568-16575.

LOUTH, J. D., 2018. *The changing face of marketing* [online]. [cit. 5. januára 2021]. Dostupné na internete: https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-changing-face-of-marketing#.

MA, L., SUN, B., 2020. Machine learning and AI in marketing – Connecting computing power to human insights. *International Journal of Research in Marketing*, 37(3), 481-504.

MATHWORKS, 2021. *Improving Classification Trees and Regression Trees* [online]. [cit. 12. decembra 2021]. Dostupné na internete: https://www.mathworks.com/help/stats/improving-classification-trees-and-regression-trees.html.

MCARDLE, J. J., RITSCHARD, G., 2014. *Contemporary Issues in Exploratory Data Mining in Behavioral Sciences*. New York, NY: Routeledge, 496 s. ISBN 978-04-1581-709-7.

MCCARTY, J. A., HASTAK, M., 2007. Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research,* 60(6), 656-662.

MINDSHARE SLOVAKIA, 2016. *Programmatic – postrach alebo príležitosť?* [online]. Brightminds by Mindshare Slovakia [cit. 16. decembra 2020]. Dostupné na internete: https://blog.mindshare.sk/2016/11/07/digital/programmatic-postrach-prilezitost/.

MINDSHARE SLOVAKIA, 2017. *Cieľte efektívnejšie vďaka programmaticu!* [online]. Brightminds by Mindshare Slovakia [cit. 16. decembra 2020]. Dostupné na internete: https://blog.mindshare.sk/2017/02/27/digital/cielte-efektivnejsie-vdaka-programmaticu/.

MINDSHARE SLOVAKIA, 2018. *Sprievodca skratkami v programmaticu – časť 1 (DSP, SSP, AD EXCHANGE)* [online]. Brightminds by Mindshare Slovakia [cit. 16. decembra 2020]. Dostupné na internete: https://blog.mindshare.sk/2018/04/16/digital/sprievodca-skratkami-v-programmaticu-cast-1-dsp-ssp-ad-exchange/.

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.

MOHLER, T., 2020. *The 7 Basic Functions of Text Analytics & Text Mining* [online]. [cit. 12. októbra 2021]. Dostupné na internete: https://www.lexalytics.com/lexablog/text-analytics-functions-explained.

MURÁŇ, J., 2019. *Algoritmy strojového učenia III. – Učenie formou odmeňovania* [online]. [cit. 15. decembra 2020]. Dostupné na internete: https://umelainteligencia.sk/algoritmy-strojoveho-ucenia-iii-ucenie-formnou-odmenovania/.

NAGPAL, A., 2017. *Decision Tree Ensembles- Bagging and Boosting* [online]. [cit. 10. januára 2021]. Dostupné na internete: https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9.

NAGWANI, N., K., 2010. Clustering Based URL Normalization Technique for Web Mining. *International Conference on Advances in Computer Engineering*, 349-351.

NETO, J., 2020. *You are correct: Machine learning is easy to understand* [online]. [cit. 4. decembra 2020]. Dostupné na internete: https://medium.com/xnewdata/you-are-correct-machine-learning-is-easy-to-understand-72950128c678.

OLSON, D. L., CHAE, B., 2012. Direct marketing decision support through predictive customer response modeling. *Decision Support Systems*, 54(1), 443-451.

OSHIRO, T. et al., 2012. How Many Trees in a Random Forest? *In: Perner P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science, vol 7376. Springer, Berlin, Heidelberg.*

POOMAGAL, S., HAMSAPRIYA, T., 2011. K-Means for Search Results Clustering Using URL and Tag Contents. *International Conference on Process Automation, Control and Computing*, 1-7.

QIAN, B. et al., 2019. *Orchestrating the Development Lifecycle of Machine Learning-Based IoT Applications: A Taxonomy and Survey* [online]. [cit. 10. decembra 2020]. Dostupné na internete: https://www.researchgate.net/publication/336642133_Orchestrating_the_Development_Lifecycle_of_Machine_Learning-Based_IoT_Applications_A_Taxonomy_and_Survey.

REIS, P. C. A., PATETTA, M., 2021. *Introduction to Statistical and Machine Learning Methods for Data Science*. Cary, NC: SAS Institute Inc., 170 s. ISBN 978-1-953329-64-6.

SAS SUPPORT, 2012. *Getting Started with SAS Text Miner 12.1*. [online]. [cit. 22. októbra 2021]. Dostupné na internete: https://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf.

SCIKIT-LEARN, 2020. *1.10. Decision Trees* [online]. [cit. 6. januára 2021]. Dostupné na internete: https://scikit-learn.org/stable/modules/tree.html.

SONG, Y., LU, Y., 2015. Decision tree methods: Applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130-135.

STATISTA, 2021. *Internet advertising spending worldwide from 2007 to 2022, by format* [online]. [cit. 2. februára 2021]. Dostupné na internete: https://www.statista.com/statistics/276671/global-internet-advertising-expenditure-by-type/.

STATISTA, 2022. *Internet users in the world 2022* [online]. [cit. 28. apríla 2022]. Dostupné na internete: https://www.statista.com/statistics/617136/digital-population-worldwide/

STECANELLA, B., 2019. *Understanding TF-ID: A Simple Introduction* [online]. [cit. 24. októbra 2021]. Dostupné na internete: https://monkeylearn.com/blog/what-is-tf-idf/.

SYAM, N., KAUL, R., 2021. *Machine Learning and Artificial Intelligence in Marketing and Sales*. Emerald Publishing Limited, 284 s. ISBN 978-18-0043-881-1.

ŠOLTÉS, E., 2019. *Regresná a korelačná analýza s aplikáciami v softvéri SAS*. 1. vyd. Bratislava: Letra Edu, 238 s. ISBN 978-80-89962-38-9.

ŠOLTÉS, E., TÁBORECKÁ-PETROVIČOVÁ, J., ŠIPOLDOVÁ, R., 2020. Targeting of Online Advertising Using Logistic Regression. *E+M. Ekonomie a management: vědecký ekonomický časopis,* 23(4), 197-214.

ŠOLTÉS, E., VOJTKOVÁ, M., ŠOLTÉSOVÁ, T., 2018. Work Intensity of Households: Multinomial Logit Analysis and Correspondence Analysis for Slovak Republic. *Statistika: statistics and economy journal*, 98(1), 19-36.

TEREK, M., HORNÍKOVÁ, A., LABUDOVÁ, V., 2010. *Hĺbková analýza údajov*. 1. vyd. Bratislava: Iura Edition, 265 s. ISBN 978-80-8078-336-5.

VOJTKOVÁ, M., STANKOVIČOVÁ, I., 2020. *Viacrozmerné štatistické metódy s aplikáciami v softvéri SAS*. 2. vyd. Bratislava: Letra Edu, 320 s. ISBN 978-80-89962-58-7.

WARD, J. H., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244.

WICKLIN, R., 2017. *The average bootstrap sample omits 36.8% of the data* [online]. [cit. 21. decembra 2021]. Dostupné na internete: https://blogs.sas.com/content/iml/2017/06/28/average-bootstrap-sample-omits-data.html.

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).

YAGCIOGLU, S., 2020. *Classical Examples of Supervised vs. Unsupervised Learning in Machine Learning* [online]. [cit. 7. decembra 2020]. Dostupné na internete: https://www.springboard.com/blog/lp-machine-learning-unsupervised-learning-supervised-learning/.

YIU, T., 2019. *Understanding Random Forest* [online]. [cit. 10. januára 2021]. Dostupné na internete: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

ZAWACKI, T., 2019. *Why Consumers Prefer Personalization* [online]. [cit. 14. decembra 2020]. Dostupné na internete: https://multichannelmerchant.com/blog/why-consumers-prefer-personalization/.

ZHANG, M. et al., 2015. A sampling method based on URL clustering for fast web accessibility evaluation. *Frontiers Inf Technol Electronic Eng*, 16, 449–456.