SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Tea Radić, Vilim Trakoštanec, Mišo Zagorec

Online Shoppers Purchasing Intention Dataset

Varaždin, 2025.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Tea Radić

Vilim Trakoštanec

Mišo Zagorec

Studij: Diplomski studij informatike

Online Shoppers Purchasing Intention Dataset

PROJEKT

Mentor/Mentorica:

Izv. prof. dr. sc. Dijana Oreški

Varaždin, studeni 2025.

Sadržaj

1. Razumijevanje domene

Kako bi smo razumjeli domenu odabranih podataka, na početku projekta proanalizirat ćemo 4 druga znanstvena rada koji se bave sličnom tematikom i analizom sličnih podataka. Ujedno je cilj odgovoriti na sljedeća pitanja:

- Koji su se podaci koristili u prethodnim istraživanjima na ovu temu?
- Koje metode strojnog učenja su primijenjene?
- Koje su evaluacijske metrike korištene i rezultati dobiveni u tim istraživanjima?
- Kakva je kvaliteta razvijenih modela i/ili inteligentnih sustava i koje smjernice autori preporučuju za buduća istraživanja?

1.1. Obrada drugih radova

1) Predicting E-commerce Purchase Behavior using a DQN-Inspired Deep Learning Model for enhanced adaptability (Jain, 2025)

Ovaj rad predlaže model nadahnut **Deep Q-Network** (DQN) pristupom koji u nadziranom okruženju kombinira **LSTM** slojeve s elementima iz pojačanog učenja (reinforcement learning) (npr. experience replay i epsilon-greedy) kako bi bolje uhvatio sekvencijalne obrasce ponašanja kupaca. Autori treniraju i testiraju na velikom e-commerce skupu od oko 885 tisuća sesija i svaka sesija ima 1.114 značajki. Usporedna analiza pokazuje prednost DQN-LSTM pristupa nad “klasičnim” ML modelima i standardnim dubokim mrežama u detekciji kompleksnih vremenskih uzoraka. Naglašena je skalabilnost rješenja i mogućnost primjene u stvarnim okruženjima s visoko dimenzionalnim, sekvencijskim podacima. Rad pokazuje da kombinacija metoda iz učenja pojačanjem i dubokog učenja može poboljšati predviđanje namjere kupnje i potražnje [1].

Podaci korišteni u istraživanju:

Korišteni su sekvencijski e-commerce logovi agregirani na razinu sesije (redoslijed događaja: view/cart/purchase), skup je vremenski bogat (2019.–2020.) i izrazito neuravnotežen po klasama (manje kupnji, više nekupnji), a neuravnoteženost je adresirana ponderiranjem klasa uz rani prekid treniranja (early stopping) i Adam optimizator [1].

Metode strojnog učenja primijenjene:

LSTM (Long short-term memory) model inspiriran **DQN** konceptima (experience

replay, epsilon-greedy) u nadziranom učenju koji je uspoređen s “klasičnim” ML i standardnim DL pristupima [1].

Evaluacijske metrike korištene i rezultati:

Primarne metrike: **točnost**, **ROC-AUC**, te **precision, recall i F1** zbog neuravnoteženosti. Prijavljeno je **oko 0,88** za točnost, a za ROC-AUC oko 0,63 te bolji učinak od baselineova [1].

Kvaliteta modela i smjernice za buduća istraživanja:

Model je skalabilan i dobro hvata vremenske obrasce u sesijama. Predlaže se daljnja analiza po kategorijama i real-time implementacije[1].

2) Understanding Online Shoppers’ Purchase Intentions using Data Analytics (Eudoxus Press, 2024)

Rad se bavi time kako iz ponašanja posjetitelja web-trgovine procijeniti hoće li netko završiti kupnju. Autori su uzeli javno dostupne podatke o sesijama posjetitelja i podijelili ih na one koji donose prihod i one koji ne donose. Ideja je izgraditi više modela, usporediti ih i odabrati rješenje koje najpouzdanije razlikuje “kupce” od “razgledavača”. Posebnu pažnju posvećuju činjenici da je kupnji razmjerno malo, pa su metode prilagođene toj neravnoteži. Na kraju nude i praktične preporuke: koje stranice treba optimizirati, kada pojačati promocije i na koje skupine posjetitelja se isplati fokusirati. Poruka rada je da se iz jednostavnih mjerenja ponašanja može izvući dosta korisnih smjernica za povećanje konverzija [2].

Podaci korišteni u istraživanju:

Korišten je UCI Online Shoppers Purchasing Intention (12.330 sesija; oko 15% pozitivne klase). Značajke uključuju broj i trajanja pregleda različitih tipova stranica (Administrative/Informational/ProductRelated), BounceRates/ExitRates, PageValues, SpecialDay te kategorijske varijable (OperatingSystems, Browser, Region, TrafficType, VisitorType, Month). Provedene su imputacije, one-hot kodiranje i provjera dimenzionalnosti putem PCA (20 do 25 značajki objašnjava >94 % varijance) [2].

Metode strojnog učenja primijenjene:

Evaluirani su LR, Naive Bayes, KNN, SVM, Decision Tree, Random Forest, AdaBoost, Gradient Boosting, Bagging Tree, XGBoost, Voting i Stacking (RF/DT/XGB kao baze, LR kao meta-model). Neravnoteža klase tretirana je s Random oversampling, SMOTE i SMOTE-ENN; odabir značajki rađen je kombinacijom Information Gain/Mutual Information, Fisher score, RFE (RF/XGB), “feature shuffling” i hibridnih XGBoost postupaka [2].

Evaluacijske metrike korištene i rezultati:

Koriste se Accuracy, Precision, Recall, F1, ROC-AUC uz K-fold CV; najbolji učinci pripisuju se XGBoost/Voting (uz oprez na overfitting). Dodatno, nakon random naduzorkovanja zabilježen je nesklad vrlo visokih CV-F1 vrijednosti za Voting (npr. CV-F1 ≈ 0,96) u odnosu na nižu test točnost (oko ~0,80), što autorima služi kao signal mogućeg preučenja. Najvažnije značajke: PageValues, ProductRelated_Duration, ExitRates, TrafficType [2].

Kvaliteta razvijenog modela i smjernice za buduća istraživanja:

Konačna procjena istakla je XGBoost kao model s najboljim kompromisom između opće sposobnosti generalizacije i prepoznavanja manjinske klase, dok je Voting također vrlo snažan, ali skloniji prekomjernom prilagođavanju. Autori preporučuju cost-sensitive evaluacije, oprez u vezi s mogućim propuštanjem informacija između sklopova podataka, te operativne smjernice (npr. smanjiti Administrative/Informational stranice, proširiti ProductRelated, fokus na “returning” i “new” posjetitelje u specifičnim mjesecima, ciljanje ključnih regija/OS-a/browsera, optimizirati idle sesije) [2].

3) Factors Affecting Online Search Intention and Online Purchase Intention

Fokus je empirijski ispitati kako utilitarne i hedonističke vrijednosti online pretraživanja, percipirane koristi i rizici e-kupovine te prethodno iskustvo kupnje utječu na namjeru online pretraživanja i, posredno i neposredno, na namjeru online kupnje [3].

Podaci korišteni u istraživanju:

Uzorak: 245 anketiranih s iskustvom online kupnje knjiga (anketa 17.–25. 11. 2002.), uz 222 valjana upitnika za konačnu analizu, s većinom mlađih odraslih u metropolitanskom Seulu i sa velikim iskustvom korištenja interneta i online kupnje.

Istraživanje je obuhvatilo mjerenje utilitarne i hedonističke vrijednosti pretraživanja, percipiranih koristi i rizika internetske kupnje, iskustva kupnje te namjere pretraživanja i kupnje. Za svaku varijablu korištene su skale sa više stavci, koje su prethodno prilagođene kroz preliminarne analize i pilot istraživanje na uzorku od 50 ispitanika [3].

Metode strojnog učenja primijenjene:

Strukturalno modeliranje (SEM) s AMOS 4.0, uz eksplorativnu faktorsku analizu (Varimax), Cronbach α za pouzdanost, te konfirmatornu faktorsku analizu za provjeru konvergentne i diskriminantne valjanosti.

Kriteriji prikladnosti modela: izvještavaju se χ^2, GFI, AGFI, RMSR i NFI, a diskriminantna valjanost provjerena je intervalima pouzdanosti korelacija koji ne uključuju 1.0 i značajnim standardiziranim opterećenjima [3].

Evaluacijske metrike korištene i rezultati:

Prilagođenost modela: $\chi^2 = 156.676$, GFI = 0.933, AGFI = 0.900, RMSR = 0.078, NFI = 0.930, što ukazuje na dobru prikladnost mjernog i strukturnog modela za promatrane konstrukte.

Ključne putanje: namjera pretraživanja pozitivno utječe na namjeru kupnje ($\beta 11 = 0.480$, $t = 3.536$, $p < 0.01$), utilitarna vrijednost povećava namjeru pretraživanja ($\gamma 11 = 0.296$, $t = 2.068$, $p < 0.05$), a hedonistička vrijednost ima još jači pozitivan učinak na namjeru pretraživanja ($\gamma 12 = 0.720$, $t = 2.803$, $p < 0.01$).

Dodatni učinci: percipirane koristi marginalno pozitivno djeluju na namjeru pretraživanja ($\gamma 13 = 0.149$, $t \approx 1.61$, $p < 0.1$), percipirani rizik nije značajno povezan s namjerom pretraživanja ($\gamma 14 = -0.074$, $t = -0.759$, n. s.), a prethodno online iskustvo pozitivno utječe i na namjeru pretraživanja ($\gamma 15 = 0.283$, $t = 3.111$, $p < 0.01$) i na namjeru kupnje ($\gamma 25 = 0.712$, $t = 6.911$, $p < 0.01$) [3].

Kvaliteta razvijenog modela i smjernice za buduća istraživanja:

Pouzdanost mjernih ljestvica: Cronbach α za konstrukte je iznad preporučenog praga 0.7 nakon uklanjanja stavki koje narušavaju unutarnju konzistenciju, a faktorska opterećenja i izdvojena varijanca potvrđuju adekvatnu mjernu stabilnost.

Valjanost i stabilnost: konvergentna valjanost potvrđena je značajnim standardiziranim opterećenjima, a diskriminantna valjanost time što intervali pouzdanosti korelacija među konstruktima ne uključuju 1.0, uz globalne indekse prilagodbe koji podupiru prikladnost modela [3].

Smjernice i preporuke za budući rad:

Naglasiti i hedonističke i utilitarne elemente iskustva pretraživanja na web-trgovinama, s posebnim fokusom na iskustvene i zabavne aspekte koji su pokazali jači doprinos namjeri pretraživanja. Proširiti analizu na dodatne faze procesa odlučivanja (prepoznavanje problema, evaluacija alternativa, postkupovno ponašanje), ispitati scenarije među-kanalnog ponašanja (npr. offline pretraga – online kupnja i obratno) te uključiti starije i raznolikije segmente potrošača i dodatne moderatorske varijable [3].

4) Purchase intention and purchase behavior online: A cross-cultural approach

Fokus je ispitati prethodnike namjere online kupnje i njihov utjecaj na stvarno online kupovno ponašanje, te testirati ulogu nacionalne kulture u usporedbi Kolumbije i Španjolske.

Podaci korišteni u istraživanju:

600 anketa u urbanim sredinama (Valencia, Španjolska i Cali, Kolumbija) tijekom 11. mjeseca 2014 do 2. mjeseca 2015 uz filtere za 18+ i kupnju online u posljednjih 6 mjeseci. Valjano N=585 (Kolumbija 291, Španjolska 294) uz etičko odobrenje CESA i obradu u EQS 6.3 i SmartPLS [4].

Kolumbija većinom žene (53%), 18–39 god. 84%, raznolika internetska staž. Španjolska uravnotežen spol, gotovo svi neudan/neoženjen, svi <39 god., s većim udjelom >7 godina iskustva na internetu [4].

Metode strojnog učenja primijenjene:

CFA za validaciju mjernog modela na ukupnom uzorku i zasebno po zemljama, s procjenom kompozitne pouzdanosti i AVE te Fornell–Larcker kriterijem diskriminantne valjanosti [4].

Test invarijantnosti mjernog instrumenta između zemalja (equal form, equal loadings) putem ΔCFI kriterija Cheung–Rensvold, uz potvrđenu invarijantnost (ΔCFI=0.006).

SEM i multi-grupna analiza s bootstrapom i PLS-SEM za provjeru hipoteza i kulturne moderacije, uz kontrolu dobi, spola, prihoda i korisničkog iskustva [4].

Evaluacijske metrike korištene i rezultati:

Prilagodba mjernog modela: NFI=0.925, NNFI=0.928, CFI=0.945, RMSEA=0.065, SRMR=0.055 za ukupni uzorak, uz CR 0.75–0.93 i AVE >0.6, što potvrđuje dobru konstruktnu pouzdanost i valjanost.

Invarijantnost mjerenja: jednake forme i opterećenja održive su između Kolumbije i Španjolske (ΔCFI=0.006), što omogućuje smisleno uspoređivanje putnih koeficijenata među skupinama.

Kontrole: korisničko iskustvo pozitivno i značajno predviđa učestalost online kupnje u ukupnom uzorku, dok je prihod različito povezan s ponašanjem u Španjolskoj (negativno značajno) u odnosu na Kolumbiju (nesignifikantno), s razlikom putova p=0.017.

Ključne putanje i moderacije:

- Namjera → ponašanje: u Kolumbiji nesignifikantno, u Španjolskoj negativno i značajno, potvrđujući kulturnu moderaciju i diskontinuitet između namjere i stvarnog ponašanja u razvijenijem uzorku.
- Stav → namjera: pozitivno i značajno u obje zemlje, s većim koeficijentom u Španjolskoj.
- PBC → namjera: pozitivno i značajno u obje zemlje, bez značajne razlike
- Samoučinkovitost → namjera: značajno pozitivno u Kolumbiji, nesignifikantno u Španjolskoj, uz značajnu moderaciju kulture.
- EOU → stav: značajno pozitivno u Španjolskoj, nesignifikantno u Kolumbiji, također s razlikom među kulturama.
- Korisnost → stav: značajno pozitivno u obje zemlje, jače u Kolumbiji, s potvrđenom moderacijom ($p \approx 0.04$).
- Impuls → namjera: pozitivno značajno u Kolumbiji, nesignifikantno negativno u Španjolskoj, s kulturnom moderacijom.
- EOU → impuls: nesignifikantno u obje zemlje, ali smjer pozitivan u Kolumbiji i negativan u Španjolskoj uz značajnu razliku putova (moderacija).
- Kompatibilnost → namjera: značajno pozitivno u obje zemlje, snažnije u Kolumbiji, s naznakom kulturne razlike.
- Subjektivne norme i PIIT → namjera: u obje zemlje nesignifikantno [4].

Kvaliteta razvijenog modela i smjernice za buduća istraživanja:

Pouzdanost i valjanost: kompozitne pouzdanosti iznad 0.7 i AVE iznad 0.5 za većinu konstrukata, uz zadovoljene kriterije konvergentne i diskriminantne valjanosti, što podupire kvalitetu mjernog dijela SEM-a. Potvrđena metrijska invarijantnost koja implicira da se uočene razlike u putnim koeficijentima mogu pripisati stvarnim razlikama u odnosima, a ne mjernim artefaktima [4].

Smjernice i preporuke za budući rad:

Mjerenje kulture na individualnoj razini, umjesto oslanjanja na nacionalne indekse, empirijski mjeriti kulturne vrijednosti svakog ispitanika radi finije analize moderacije. Proširenje mjerenja ponašanja, uz učestalost kupnje uključiti prosječan iznos i udio online budžeta kako bi se preciznije modelirala veza namjera–ponašanje. Uključivanje vanjskih čimbenika, dakle ekonomski kontekst, tržišne šokove i druge situacijske varijable koji mogu razdvojiti namjere od ponašanja, osobito u razvijenim tržištima [4].

2. Razumijevanje podataka

2.1. Opis skupa podataka

Skup podataka sastoji se od podataka o online kupcima prikupljenih s web trgovine. Cilj je predvidjeti hoće li korisnik izvršiti kupovinu. Skup sadrži 18 varijabli, od kojih su neke numeričke, a neke kategorijske.

- **Administrative** - broj stranica administrativnog tipa koje je korisnik posjetio.
- **Administrative_Duration** - ukupno vrijeme provedeno na tim stranicama.
- **Informational** - broj informativnih stranica posjećenih tijekom sesije.
- **Informational_Duration** - vrijeme provedeno na informativnim stranicama.
- **ProductRelated** - broj stranica povezanih s proizvodima koje je korisnik pregledao.
- **ProductRelated_Duration** - ukupno vrijeme provedeno na stranicama proizvoda.
- **BounceRates** - stopa korisnika koji napuste stranicu bez daljnje interakcije.
- **ExitRates** - stopa izlaska sa stranica.
- **PageValues** - procijenjena vrijednost pojedine stranice za mogućnost konverzije.
- **SpecialDay** - pokazuje odnos prema posebnim danima (npr. blagdani, akcije).
- **Month** - mjesec u kojem je sesija ostvarena.
- **OperatingSystems** - operacijski sustav korisnika.
- **Browser** - korišteni web preglednik.
- **Region** - geografska regija korisnika.
- **TrafficType** - tip izvora prometa (npr. direktan, oglas, pretraga).
- **VisitorType** - vrsta posjetitelja (novi ili ponovni).
- **Weekend** - označava je li sesija bila tijekom vikenda.
- **Revenue** - ciljana varijabla (TRUE/FALSE) koja pokazuje je li kupnja ostvarena.

2.2. Prikaz deskriptivnih statistika i distribucija varijabli

Numeričke varijable

	count	mean	std	min	25%	median	75%	max
Administrative	12330.0	2.315166	3.321784	0.0	0.0	1.0	4.0	27.0
Administrative_Duration	12330.0	80.818611	176.779107	0.0	0.0	7.5	93.25625	3398.75
Informational	12330.0	0.503569	1.270156	0.0	0.0	0.0	0.0	24.0
Informational_Duration	12330.0	34.472398	140.749294	0.0	0.0	0.0	0.0	2549.375
ProductRelated	12330.0	31.731468	44.475503	0.0	7.0	18.0	38.0	705.0
ProductRelated_Duration	12330.0	1194.74622	1913.669288	0.0	184.1375	598.936905	1464.157214	63973.52223
BounceRates	12330.0	0.022191	0.048488	0.0	0.0	0.003112	0.016813	0.2
ExitRates	12330.0	0.043073	0.048597	0.0	0.014286	0.025156	0.05	0.2
PageValues	12330.0	5.889258	18.568437	0.0	0.0	0.0	0.0	361.763742
SpecialDay	12330.0	0.061427	0.198917	0.0	0.0	0.0	0.0	1.0
OperatingSystems	12330.0	2.124006	0.911325	1.0	2.0	2.0	3.0	8.0
Browser	12330.0	2.357097	1.717277	1.0	2.0	2.0	2.0	13.0
Region	12330.0	3.147364	2.401591	1.0	1.0	3.0	4.0	9.0
TrafficType	12330.0	4.069586	4.025169	1.0	2.0	2.0	4.0	20.0

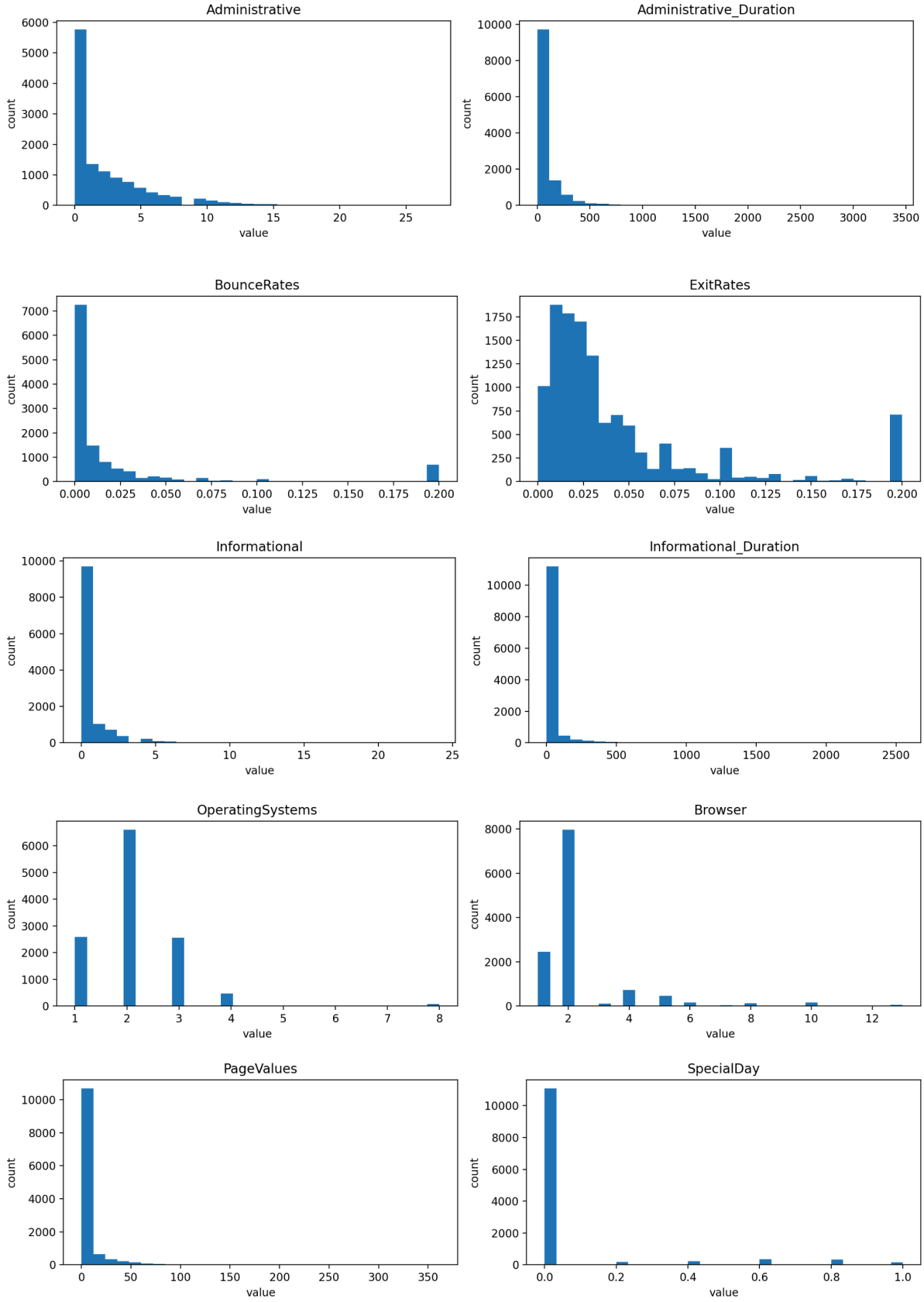
Slika 1: Deskriptivna statistika numeričkih varijabli

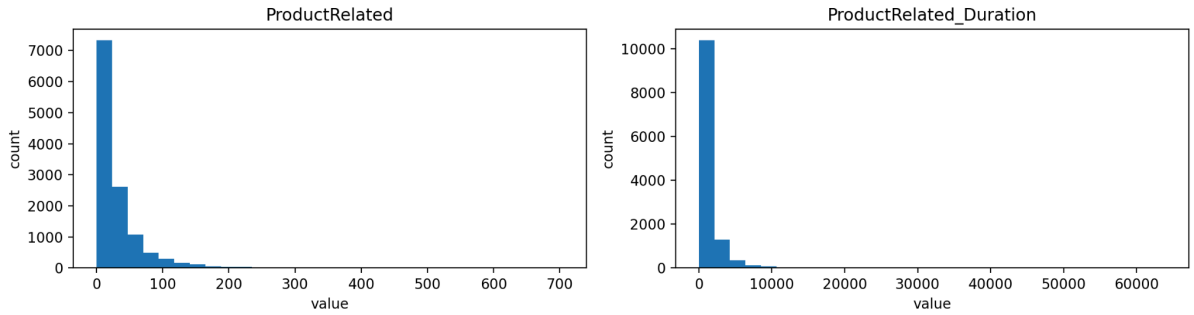
Deskriptivna statistika

```
=== Deskriptivna statistika (kategorijske) ===

              unique                mode  mode_freq
variable
Month             10                 May       3364
VisitorType        3   Returning_Visitor      10551
Weekend            2               False       9462
Revenue            2               False      10422
```

Slika 2: Deskriptivna statistika kategorijskih varijabli

Histogrami numeričkih varijabli





Slika 3-8: Histogrami numeričkih varijabli

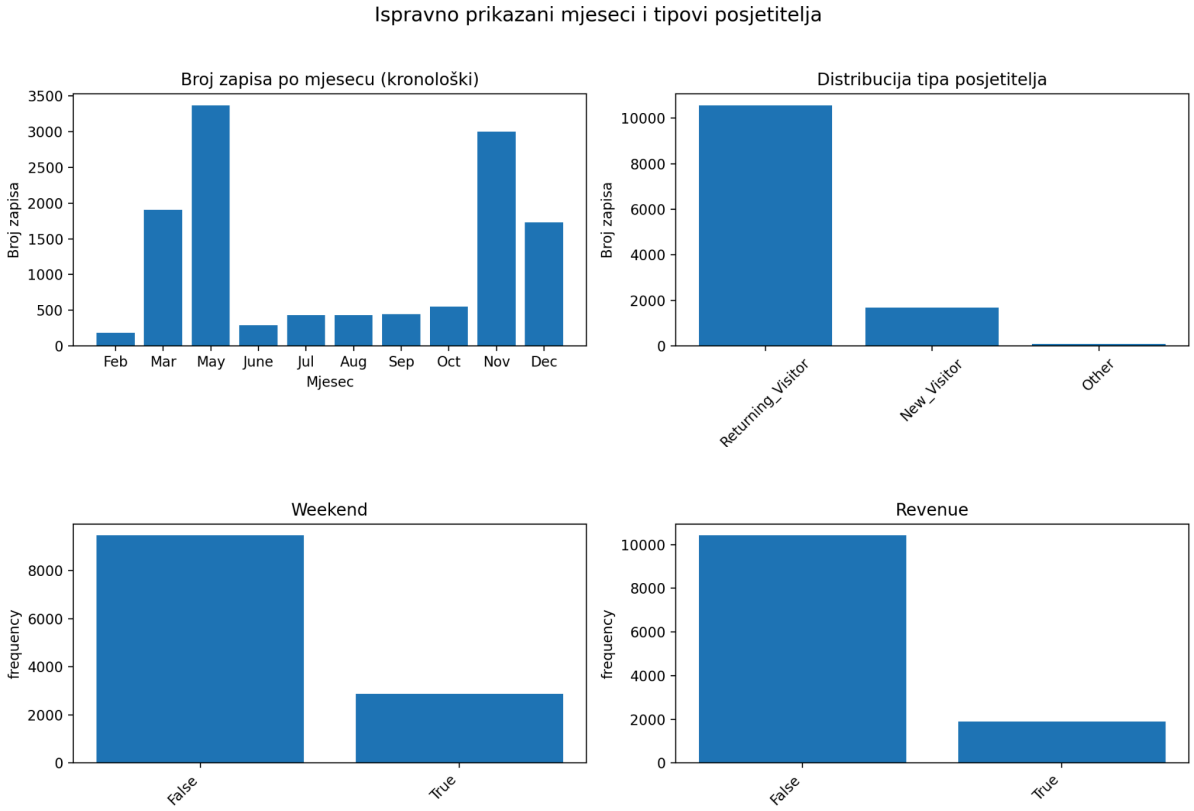
Distribucija većine numeričkih varijabli pokazuje izraženu pozitivnu asimetriju (right-skewness), što znači da se najveći broj vrijednosti koncentrira uz donju granicu, dok se rjeđe pojavljuju visoko ekstremne vrijednosti.
Promatrajući varijable kao što su Administrative_Duration, Informational_Duration, ProductRelated_Duration, PageValues te BounceRates i ExitRates, vidljivo je da većina korisnika provodi relativno kratko vrijeme pregledavajući stranice, dok manji broj korisnika provodi znatno više vremena, što rezultira dugim repovima distribucije.

Varijable poput Region, TrafficType, OperatingSystems i Browser iako su numeričke, predstavljaju kategorijske kodirane vrijednosti, pa njihovi histogrami izgledaju kao diskretne raspodjele s jasno izdvojenim najčešćim vrijednostima.

Ukratko:

- većina numeričkih varijabli koncentrirana je na niže vrijednosti,
- prisutni su outlieri, što je očekivano u web prometu,
- neke numeričke varijable zapravo predstavljaju kategoričke kodove.

Distribucija kategorijskih varijabli



Slika 9-10: Histogrami kategorijskih varijabli

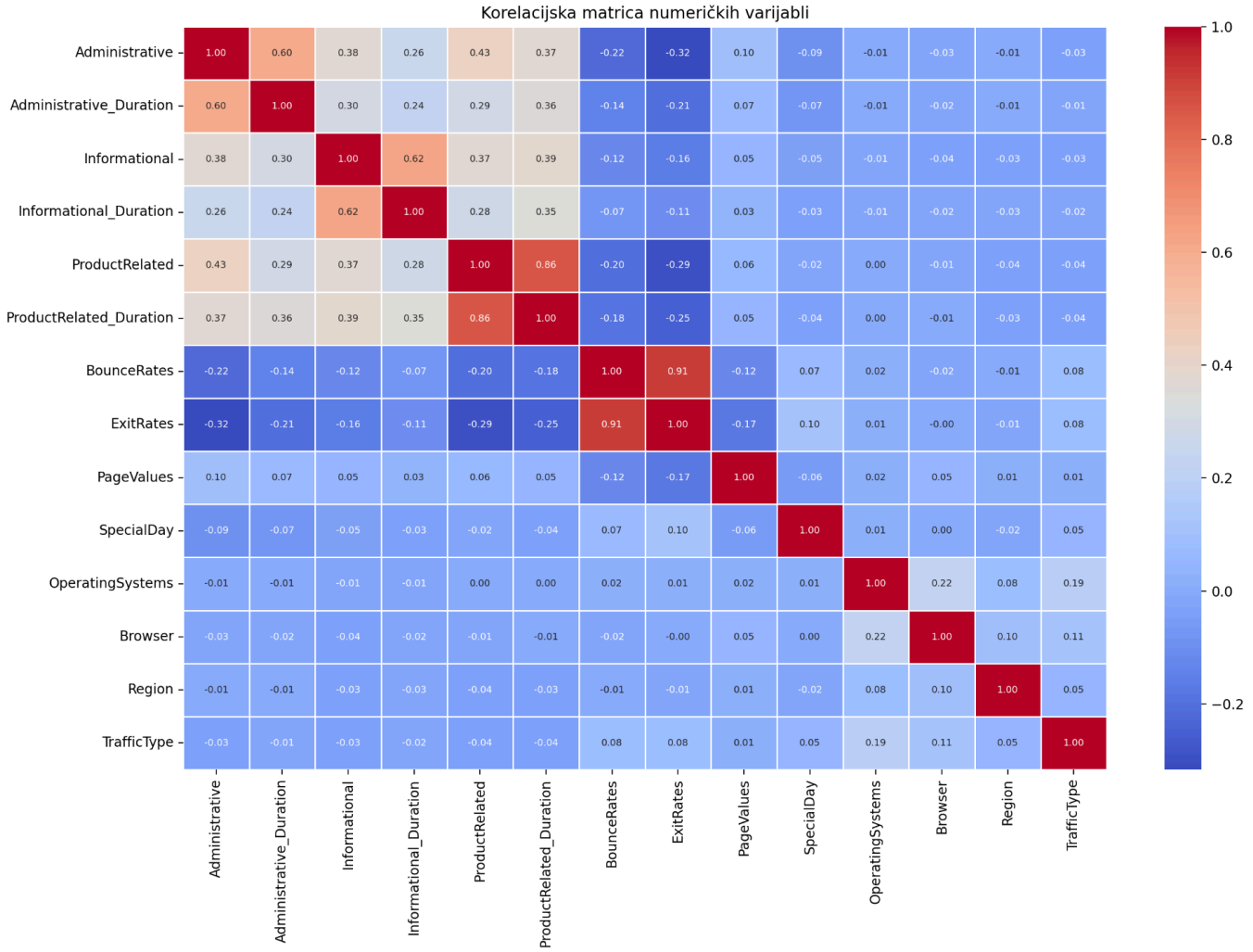
Kod kategorijskih varijabli primjećuje se dominantna prisutnost pojedinih kategorija. Varijabla VisitorType pokazuje da je većina posjetitelja Returning_Visitor, što sugerira da se kupci vraćaju na stranicu više puta. Varijabla Weekend i Month pokazuju sezonalnost i promjene u aktivnosti, pri čemu se najveći broj sesija javlja u svibnju i studenom, dok je manji broj posjetitelja aktivan tijekom drugih mjeseci.

Varijabla Revenue također pokazuje neravnotežu klasa, jer je broj sesija koje nisu rezultirale kupnjom znatno veći od onih koje jesu.

Ukratko:

- često postoji jedna dominantna kategorija (npr. Returning_Visitor),
- distribucije pokazuju sezonske obrasce (npr. Month),
- ciljna varijabla Revenue je neizbalansirana, što je važna napomena za modeliranje.

2.3. Odnos između varijabli



Slika 11: Matrica korelacija

Najveće pozitivne korelacije u podacima pojavljuju se između varijabli koje mjere broj posjećenih stranica i vrijeme provedeno na njima. To je očekivano jer se radi o direktno povezanim korisničkim radnjama.

- ProductRelated – ProductRelated_Duration (r = 0.86)
  Što više proizvoda korisnik pregledava, to duže provodi vremena na tim stranicama – potpuno očekivano ponašanje.

- BounceRates -- ExitRates (r = 0.91)
  Sesije s visokom stopom odbijanja najčešće završavaju brzim izlaskom s web stranice.

- Administrative – Administrative_Duration (r = 0.60)
  Što više administrativnih stranica korisnik otvara, to više vremena provodi na njima.

- Informational – Informational_Duration (r = 0.62)
  Isto vrijedi i za informacijske stranice, više posjeta vodi do više provedenog vremena.

Iako su negativne korelacije relativno slabe, one ipak ukazuju na neke važne obrasce ponašanja.

- BounceRates – PageValues (r = -0.12)
  Što je korisnik skloniji napustiti stranicu odmah, to je manja vjerojatnost da će sadržaj imati vrijednost u smislu konverzije.

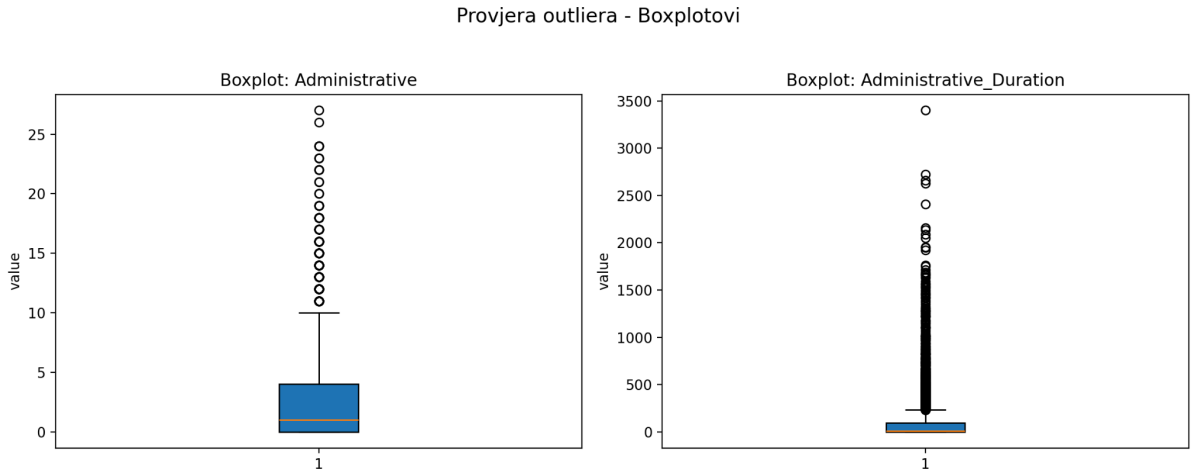
- ExitRates – PageValues (r = -0.17)
  Visoka stopa izlaza obično znači da korisnik nije nastavio prema kupovini – dakle stranica ne vodi prema željenoj akciji.

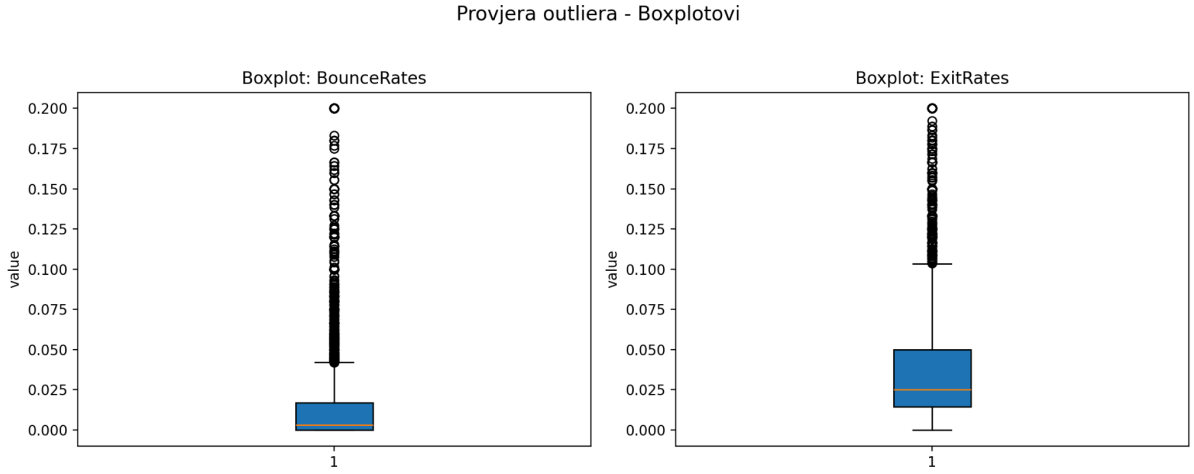
2.4 Kvaliteta podataka

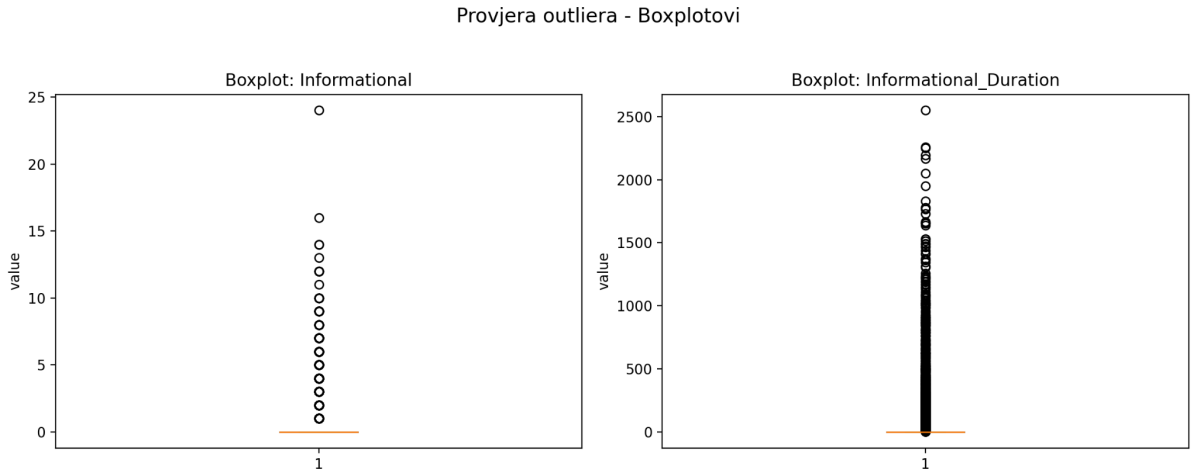
```
=== Kvaliteta podataka ===


Broj redaka: 12330
Broj stupaca: 18

Nedostajuće vrijednosti po stupcima:
Administrative             0
Administrative_Duration    0
Informational              0
Informational_Duration     0
ProductRelated             0
ProductRelated_Duration    0
BounceRates                0
ExitRates                  0
PageValues                 0
SpecialDay                 0
Month                      0
OperatingSystems           0
Browser                    0
Region                     0
TrafficType                0
VisitorType                0
Weekend                    0
Revenue                    0

Broj duplikata: 125

Rows: 12330 | Columns: 18
Missing values: none
```

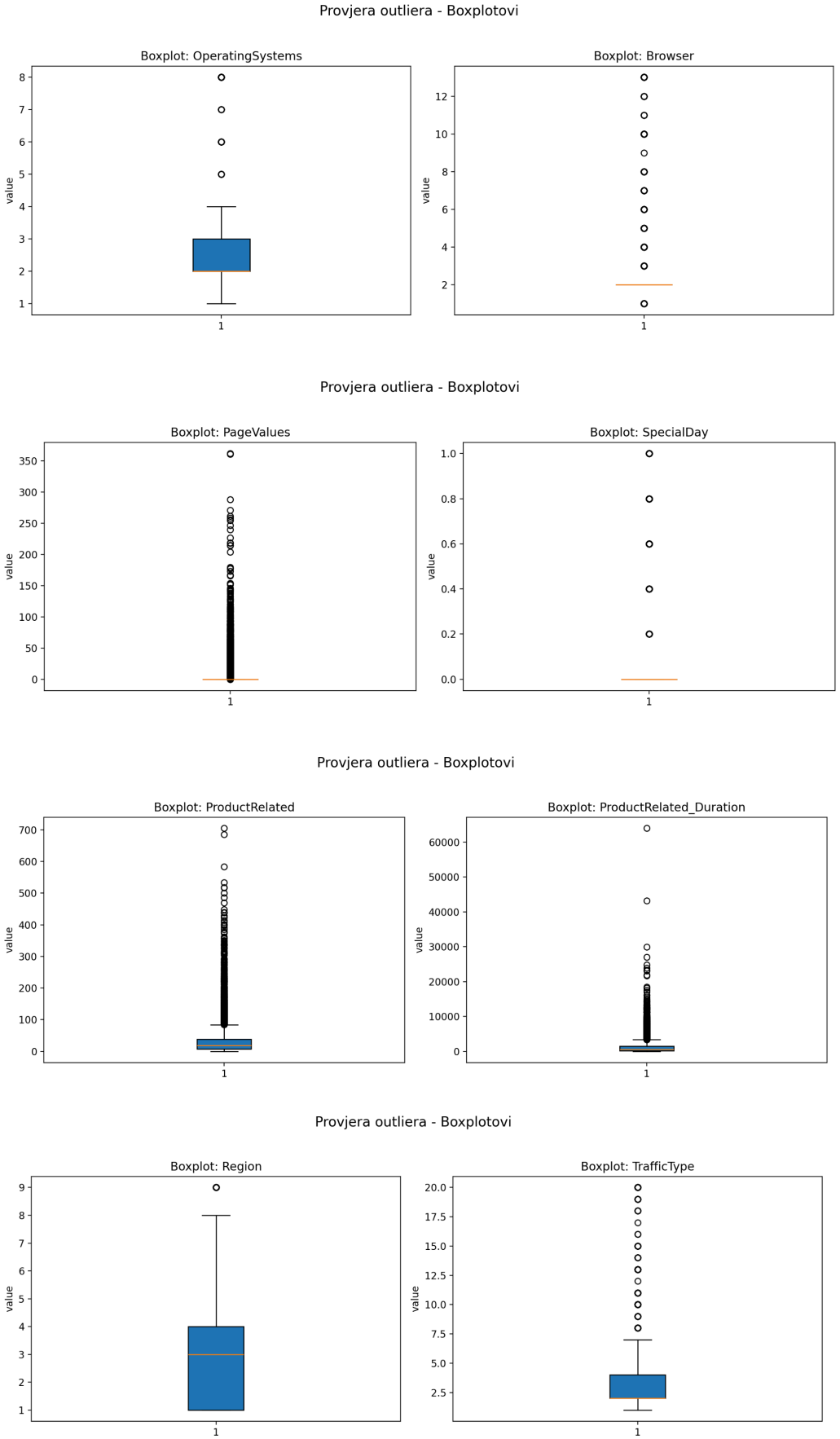
Podaci su općenito vrlo dobre kvalitete. U čitavom skupu nema nedostajućih vrijednosti, što znači da nije potrebna nikakva imputacija ili čišćenje zbog praznih zapisa. Jedina primjetna nepravilnost je postojanje otprilike 125 duplikata, što čini oko jedan posto ukupnog broja redaka. Iako to nije velika brojka, preporučljivo je ukloniti duplikate kako bi se izbjeglo dvostruko računanje i iskrivljavanje statistika ili modela strojnog učenja. Osim toga, skup podataka je uredno strukturiran i spreman za analizu.

Slika 12: Kvaliteta podataka


Provjera outliera - Boxplotovi
Boxplot: Administrative
Boxplot: Administrative_Duration
value
0
5
10
15
20
25
500
1000
1500
2000
2500
3000
3500
1


Provjera outliera - Boxplotovi
Boxplot: BounceRates
Boxplot: ExitRates
value
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
1


Provjera outliera - Boxplotovi
Boxplot: Informational
Boxplot: Informational_Duration
value
0
5
10
15
20
25
500
1000
1500
2000
2500
1



Slika 13-19: Box plot dijagrami numeričkih varijabli

Pregled boxplotova pokazuje da se u mnogim varijablama pojavljuju izraženi outlieri. To se osobito vidi u varijablama vezanim uz trajanje posjeta, kao što su *ProductRelated_Duration*, *Informational_Duration* i *Administrative_Duration*. Takvi outlieri nisu nužno greške ili "loši podaci", nego odražavaju stvarno ponašanje korisnika: manji broj njih provodi neuobičajeno puno vremena na određenim tipovima stranica. To je sasvim uobičajeno u web analitici, gdje se ponašanje korisnika jako razlikuje.

Slično se primjećuje i kod varijable *PageValues*, gdje se vidi mali broj sesija koje imaju iznimno visoku vrijednost stranica. To sugerira postojanje manjeg segmenta korisnika s visokom namjerom kupnje. Drugim riječima, iako su ti slučajevi rijetki, oni nose komercijalno vrlo važnu informaciju i ne bi ih trebalo automatski uklanjati.

S druge strane, varijable poput *BounceRates* i *ExitRates* pokazuju manje varijacije, no vidljive su razlike među korisnicima koji brzo napuštaju stranicu i onima koji pregledavaju više sadržaja. To također predstavlja realnu razliku u ponašanju korisnika, a ne grešku u podacima.

Kod varijabli poput *Browser*, *OperatingSystems*, *Region* i *TrafficType* boxplotovi ne ukazuju na problematične izbijajuće vrijednosti, već samo pokazuju da su neke kategorije mnogo rjeđe od drugih, što je također očekivano.

Iako se u podacima pojavljuje veliki broj outliera, oni nisu rezultat pogreške već prirodnog raspona ponašanja korisnika u on-line okruženju. Iz tog razloga **outliere ne treba uklanjati**, već ih je potrebno pravilno obraditi pri modeliranju, primjerice korištenjem log-transformacija ili robustnih metoda skaliranja. Dataset je kvalitetan, potpun i informativan te spreman za napredniju analizu ili pripremu modela.

3. Deskriptivno modeliranje (klasteriranje)

3.1. Priprema podataka

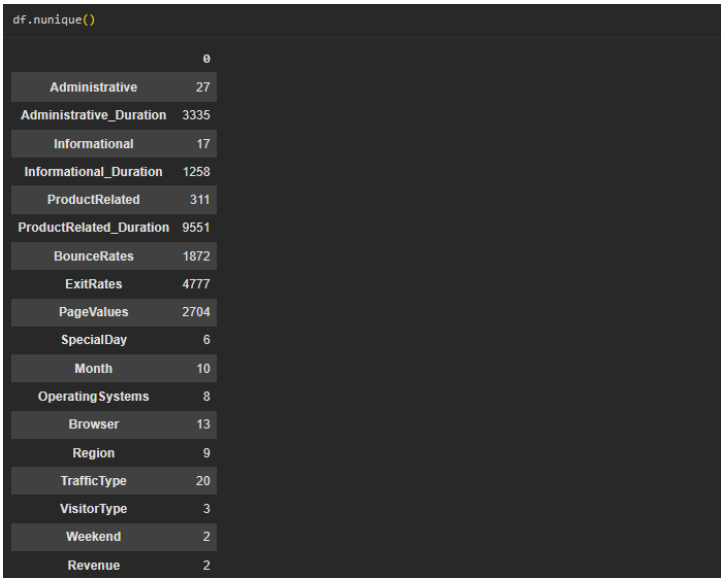
U početnom skupu podataka nalazilo se 12 330 redaka. Provjerom broja dupliciranih zapisa pomoću naredbe df.duplicated().sum() utvrđeno je da postoji 125 potpuno istih redaka (slika 20). Kako bi se izbjeglo ponavljanje podataka i mogući utjecaj na analizu, duplicirani redci uklonjeni su naredbom df.drop_duplicates(keep='first', inplace=True), koja zadržava prvo pojavljivanje zapisa. Nakon uklanjanja ponovno je izvršena provjera naredbom df.duplicated().sum(), pri čemu je dobiveno 0, što znači da u skupu više nema duplikata. Završnom provjerom broja redaka naredbom len(df) utvrđeno je da skup sada sadrži 12 205 zapisa i kao takav je spreman za daljnju obradu.

```
import pandas as pd

df = pd.read_csv("online_shoppers_intention.csv")

print("Broj duplih redaka prije brisanja:", df.duplicated().sum())

df.drop_duplicates(keep='first', inplace=True)

print("Broj duplih redaka nakon brisanja:", df.duplicated().sum())
print("Novi broj redaka:", len(df))
```

```
Broj duplih redaka prije brisanja: 125
Broj duplih redaka nakon brisanja: 0
Novi broj redaka: 12205
```

Slika 20: Uklanjanje dupliciranih podataka

Kako bi se provjerilo postoji li stupac koji sadrži jedinstvenu vrijednost u svakom retku i koji bi mogao biti nepotreban (npr. neki ID), nad skupom podataka pokrenuta je naredba df.nunique(), što je vidljivo na slici 21. Ova naredba vraća broj različitih vrijednosti za svaki stupac. U analiziranom skupu svi stupci imaju razuman broj različitih vrijednosti i odnose se na ponašanje korisnika na webu (trajanje, broj stranica, tip posjetitelja, mjesec i sl.), pa nije uočen stupac koji bi bio samo identifikator i koji bi trebalo ukloniti. Zbog toga su svi postojeći stupci zadržani u daljnjoj analizi.



```
df.nunique()
```

	0
Administrative	27
Administrative_Duration	3335
Informational	17
Informational_Duration	1258
ProductRelated	311
ProductRelated_Duration	9551
BounceRates	1872
ExitRates	4777
PageValues	2704
SpecialDay	6
Month	10
OperatingSystems	8
Browser	13
Region	9
TrafficType	20
VisitorType	3
Weekend	2
Revenue	2

Slika 21: Provjera nepotrebnih stupaca

Provjerom strukture podataka naredbom df.info(), prikazanoj na slici 22, utvrđeno je da svi atributi imaju jednak broj zapisa (12 205) te da u skupu nema nedostajućih vrijednosti. Zbog toga imputacija nedostajućih podataka nije bila potrebna. U slučaju da su postojale NaN vrijednosti, one bi bile popunjene odgovarajućom statističkom mjerom (npr. sredinom ili medijanom za numeričke atribute, odnosno najčešćom vrijednošću za kategorijske).

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12205 entries, 0 to 12204
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12205 non-null  int64
 1   Administrative_Duration  12205 non-null  float64
 2   Informational            12205 non-null  int64
 3   Informational_Duration   12205 non-null  float64
 4   ProductRelated           12205 non-null  int64
 5   ProductRelated_Duration  12205 non-null  float64
 6   BounceRates              12205 non-null  float64
 7   ExitRates                12205 non-null  float64
 8   PageValues               12205 non-null  float64
 9   SpecialDay               12205 non-null  float64
 10  Month                    12205 non-null  object
 11  OperatingSystems         12205 non-null  int64
 12  Browser                  12205 non-null  int64
 13  Region                   12205 non-null  int64
 14  TrafficType              12205 non-null  int64
 15  VisitorType              12205 non-null  object
 16  Weekend                  12205 non-null  bool
 17  Revenue                  12205 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
```

Slika 22: Provjera nedostajućih vrijednosti

Za potrebe klasteriranja bilo je potrebno podatke dodatno prilagoditi algoritmu. Najprije su odabrani atributi koji opisuju ponašanje korisnika na web stranici, poput broja posjećenih administrativnih, informativnih i stranica s proizvodima te vremena provedenog na njima, kao i pokazatelja BounceRates, ExitRates, PageValues i SpecialDay. Atribut Revenue izdvojen je jer predstavlja ishod (je li došlo do kupnje) i ne treba sudjelovati u formiranju klastera. Budući da skup sadrži i kategorijske varijable (Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend), one su prije klasteriranja pretvorene u numerički oblik pomoću funkcije pd.get_dummies(), čime je za svaku kategoriju dobiven zaseban binarni stupac (0/1). Iako su u izvornom opisu podataka varijable poput OperatingSystems, Browser, Region i TrafficType navedene kao cjelobrojne, u stvarnosti predstavljaju samo šifre kategorija, pa bi ih k-means mogao krivo tumačiti kao “veće” ili “manje” vrijednosti. Zbog toga su i te varijable pretvorene u binarni oblik kako bi sve kategorije bile ravnopravne u računanju udaljenosti i kako bi se dobili pouzdaniji klasteri što je vidljivo na slici 23.

```
categorical_features = [
    "Month",
    "OperatingSystems",
    "Browser",
    "Region",
    "TrafficType",
    "VisitorType",
    "Weekend"
]


df_encoded = pd.get_dummies(df, columns=categorical_features, drop_first=True, dtype=int)

df_encoded.head()
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration
0	0	0.0	0	0.0	1	0.000000
1	0	0.0	0	0.0	2	64.000000
2	0	0.0	0	0.0	1	0.000000
3	0	0.0	0	0.0	2	2.666667
4	0	0.0	0	0.0	10	627.500000

5 rows × 69 columns

Slika 23: Pretvaranje kategorijskih varijabli u binarne (one-hot kodiranje)

Nakon toga su numerički atributi skalirani pomoću StandardScalera (prikazano na slici 24) kako bi sve varijable bile na usporedivoj skali i kako pojedini atribut s velikim rasponom vrijednosti ne bi dominirao u računanju udaljenosti. Na taj je način dobiven završni skup podataka bez duplikata, bez nepotrebnih stupaca i s pripremljenim numeričkim i kategoriziranim varijablama, spreman za primjenu algoritma klasteriranja.

```
numerical_features = [
    "Administrative", "Administrative_Duration",
    "Informational", "Informational_Duration",
    "ProductRelated", "ProductRelated_Duration",
    "BounceRates", "ExitRates",
    "PageValues", "SpecialDay"
]

scaler = StandardScaler()
scaled_numeric = scaler.fit_transform(df_encoded[numerical_features])

scaled_numeric_df = pd.DataFrame(scaled_numeric, columns=numerical_features)

other_features = df_encoded.drop(columns=numerical_features)

final_df = pd.concat([scaled_numeric_df, other_features], axis=1)

X = final_df.drop(columns=["Revenue"])

print("Oblik završnog skupa za klasteriranje:", X.shape)
X.head()
```

```
Oblik završnog skupa za klasteriranje: (12205, 68)
```

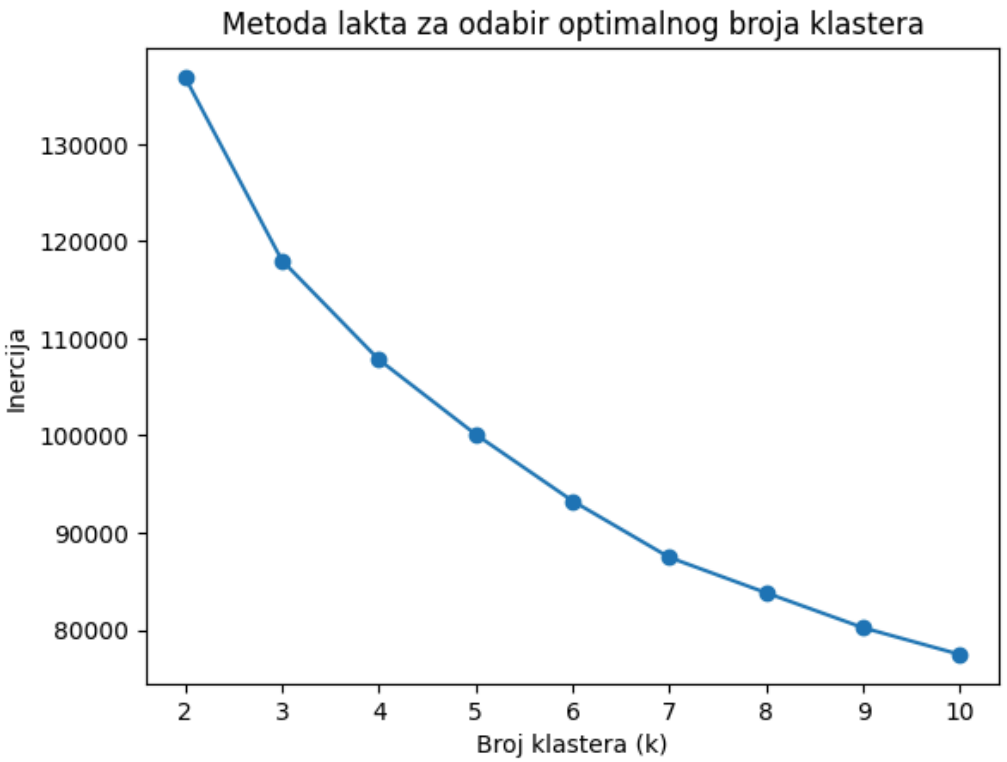
	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated
0	-0.702302	-0.460019	-0.398824	-0.246257	-0.696218
1	-0.702302	-0.460019	-0.398824	-0.246257	-0.673793
2	-0.702302	-0.460019	-0.398824	-0.246257	-0.696218
3	-0.702302	-0.460019	-0.398824	-0.246257	-0.673793
4	-0.702302	-0.460019	-0.398824	-0.246257	-0.494387

5 rows × 68 columns

Slika 24: Skaliranje numeričkih varijabli i formiranje završnog skupa za klasteriranje

3.2. Klasteriranje

Nakon pripreme podataka primijenjen je algoritam klasteriranja k-srednjih vrijednosti (K-means). Budući da ovaj algoritam traži da se unaprijed odredi broj klastera, najprije je napravljena provjera za više različitih vrijednosti k (od 2 do 10). Za svaku vrijednost izračunata je inercija i prikazana metodom lakta. Na dobivenom grafu na slici 25. vidi se da se inercija najviše smanjuje kod manjih vrijednosti k, a zatim pad postaje sve blaži. Na temelju toga kao razumna vrijednost odabrano je k = 5, jer nakon pet klastera dodatno smanjenje inercije više nije toliko izraženo. Time je zapravo izvršena optimizacija hiperparametra, odabran je broj klastera koji dobro opisuje podatke, a istovremeno ne stvara prevelik broj grupa. Nakon odabira k model K-means je ponovno pokrenut s tom vrijednošću i svakom zapisu u skupu podataka dodijeljena je oznaka klastera.



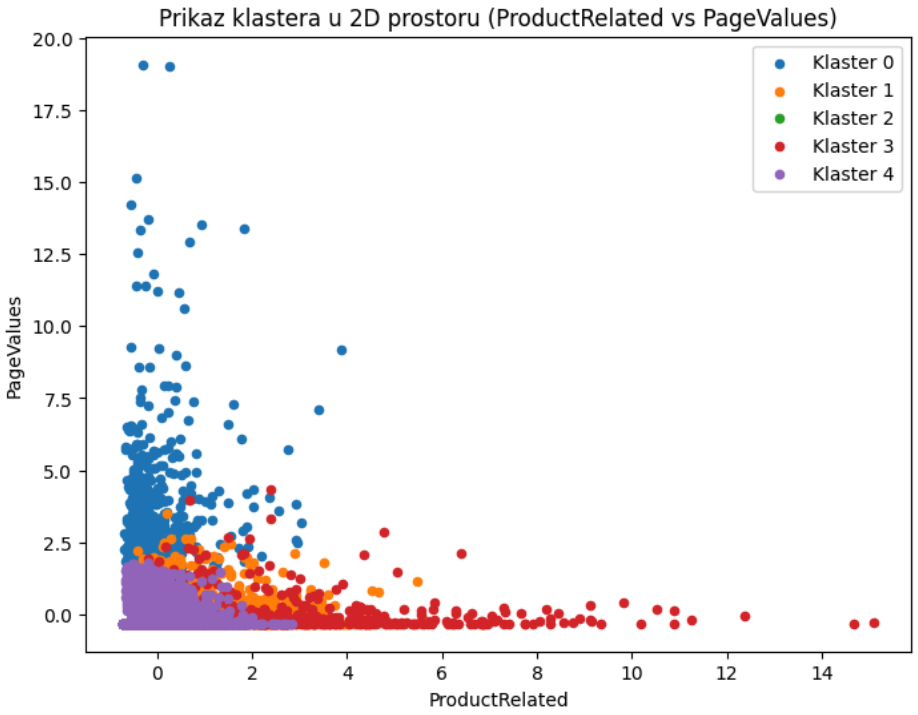
Slika 25: Metoda lakta za određivanje optimalnog broja klastera

3.3. Evaluacija

Za procjenu kvalitete dobivenih klastera izračunata je Silhouette ocjena. Za odabrani broj klastera k = 5 dobivena je vrijednost Silhouette score = 0,2109. Silhouette se kreće u rasponu od -1 do 1: vrijednosti blizu 1 označavaju jako dobro formirane i jasno odvojene klastere, vrijednosti oko 0 upućuju na to da su klasteri djelomično preklopljeni, dok negativne vrijednosti znače da su zapisi možda dodijeljeni pogrešnom klasteru. U ovom slučaju dobivena pozitivna vrijednost pokazuje da su zapisi u prosjeku bliži svom klasteru nego ostalima, ali razdvajanje nije jako izraženo. To je očekivano jer skup podataka sadrži velik broj značajki i kombinaciju numeričkih i binarnih varijabli, pa se klasteri djelomično preklapaju. Unatoč tome, rezultat potvrđuje da algoritam uspijeva prepoznati određenu strukturu u podacima i grupirati korisnike u nekoliko smislenih segmenata.

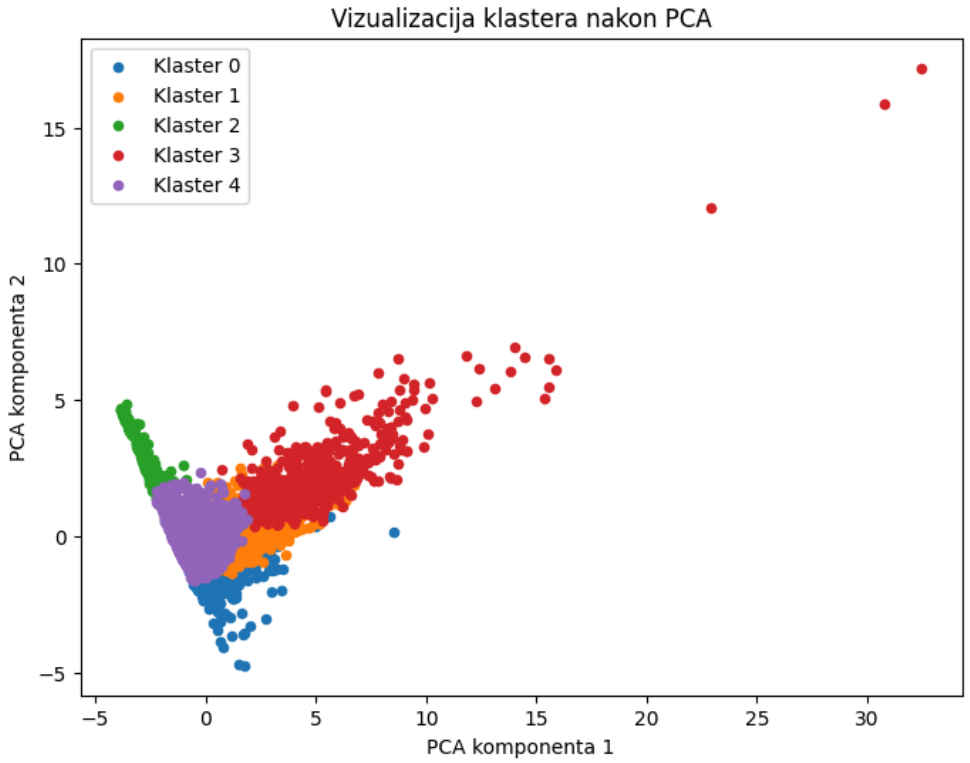
Za potrebe boljeg uvida u dobivene skupine napravljena je vizualizacija klastera prikazana na slici 26. Prvo je prikazan jednostavan 2D raspršeni dijagram u kojem su na osi x uzete vrijednosti atributa ProductRelated, a na osi y atribut PageValues, dok je svaka točka obojana prema klasteru kojem pripada. Na ovom prikazu vidi se da je velik broj posjetitelja koncentriran u području malih vrijednosti oba atributa (malo pregledanih stranica i mala ili

nikakva vrijednost košarice), pa se klasteri djelomično preklapaju što je očekivano za ovakve web-podatke.



Slika 26: Vizualizacija klastera u 2D prostoru

Kako bi se dobio pregledniji prikaz svih varijabli odjednom, napravljena je i vizualizacija klastera pomoću PCA metode (Principal Component Analysis). PCA je smanjio početni skup od nekoliko desetaka značajki na samo dvije glavne komponente koje zadržavaju najveći dio varijabilnosti podataka, pa se svaki posjetitelj sada može nacrtati kao jedna točka u 2D prostoru. Na slici 27. se vidi da se točke iste boje (istog klastera) u velikoj mjeri grupiraju na sličnim područjima, npr. ljubičasti klaster je koncentriran u užem području oko ishodišta, dok se crveni klaster proteže dijagonalno prema gore i desno, što znači da ti posjetitelji imaju sličan “smjer” ponašanja po više atributa. Uočljiva su i nekoliko točaka daleko desno koje pripadaju istom klasteru. To su vjerojatno posjetitelji s neuobičajeno visokim vrijednostima (npr. puno stranica i visoki PageValues). Takav PCA prikaz ne služi za novo klasteriranje, nego kao ilustracija da je k-means pronašao strukturu u podacima i da se skupine mogu vizualno razlikovati.



Slika 27: Vizualizacija pomoću PCA

Dobiveni rezultati klasteriranja pokazali su da se promatrani posjeti web-stranici mogu podijeliti u pet smislenih skupina koje se razlikuju prvenstveno po dubini interakcije (broj i trajanje pregledanih stranica), kvaliteti posjeta (PageValues) i ponašanju pri napuštanju stranice (Bounce/Exit).

Najveći klaster (klaster 4) čine tipični posjeti s kratkim pregledom sadržaja. Kod njih su gotovo svi numerički atributi nešto ispod prosjeka (negativne skalirane vrijednosti), što znači da korisnici otvore malo administrativnih, informativnih i produktnih stranica i relativno brzo završe sesiju. PageValues je također nizak, pa ove posjete možemo opisati kao masovnu, ali niskovrijednu grupu koja dominantno služi informiranju ili usputnom pregledavanju sadržaja.

Klaster 1 predstavlja angažiranije posjete. Kod ove skupine broj i trajanje većine vrsta stranica (administrativne, informativne, produktne) su iznad prosjeka, što znači da korisnici aktivno pretražuju stranicu i prolaze kroz više tipova sadržaja. Iako PageValues nije nužno visok, ovo su posjeti koji ozbiljnije pregledavaju ponudu i mogu predstavljati publiku bližu fazi odluke.

Posebno se izdvaja klaster 0, iako je brojčano malen. Ovdje je PageValues znatno viši nego u ostalim skupinama, a stope napuštanja su niske. To upućuje na to da se u ovoj skupini nalaze posjeti koji su doveli do neke vrijedne akcije (npr. dodavanje u košaricu ili konverzija). Ovaj klaster možemo promatrati kao najvrjedniji segment korisnika.

Klaster 2 čine posjeti izrazito visokih stopa napuštanja (BounceRates i ExitRates su daleko iznad prosjeka) uz vrlo mali broj pregledanih stranica. To je problematični segment. Korisnici dođu, vrlo brzo odustanu i ne generiraju nikakvu vrijednost. U praksi je ovaj klaster zanimljiv jer pokazuje gdje bi se mogla poboljšati relevantnost odredišnih stranica ili izvora prometa.

Napokon, klaster 3 obuhvaća najmanju, ali vrlo specifičnu skupinu korisnika koji pregledavaju puno stranica i dugo ostaju na stranici. Gotovo sve trajanja i brojevi stranica su višestruko iznad prosjeka. To su duboki, istraživački posjeti (npr. uspoređivanje proizvoda, proučavanje više kategorija). Iako im PageValues nije jednako visok kao u klasteru 0, riječ je o korisnicima s jasnim interesom i visokim potencijalom da u nekoj od sljedećih posjeta ostvare konverziju.

Klaster 0 - deskriptivna statistika:

	Administrative	Administrative_Duration	Informational	\
count	587.00	587.00	587.00	
mean	-0.01	-0.03	-0.09	
std	0.84	0.77	0.76	
min	-0.70	-0.46	-0.40	
25%	-0.70	-0.46	-0.40	
50%	-0.40	-0.26	-0.40	
75%	0.50	0.08	-0.40	
max	5.90	8.89	5.09	

	Informational_Duration	ProductRelated	ProductRelated_Duration	\
count	587.00	587.00	587.00	
mean	-0.11	-0.04	-0.03	
std	0.46	0.63	0.52	
min	-0.25	-0.70	-0.62	
25%	-0.25	-0.43	-0.37	
50%	-0.25	-0.23	-0.17	
75%	-0.25	0.11	0.15	
max	3.95	3.88	2.48	

	BounceRates	ExitRates	PageValues	SpecialDay	...	TrafficType_14	\
count	587.00	587.00	587.00	587.00	...	587.0	
mean	-0.39	-0.58	3.48	-0.25	...	0.0	
std	0.15	0.25	2.24	0.43	...	0.0	
min	-0.45	-0.90	1.52	-0.31	...	0.0	
25%	-0.45	-0.76	2.16	-0.31	...	0.0	
50%	-0.45	-0.63	2.80	-0.31	...	0.0	
75%	-0.45	-0.47	3.91	-0.31	...	0.0	
max	0.76	0.68	19.08	4.70	...	0.0	

Slika 28: Isječak deskriptivne statistike klastera

4. Odabir ciljanih varijabli

Na temelju provedenog klasteriranja i razumijevanja podataka odabrane su i ciljne varijable za drugu fazu projekta, u kojoj će se raditi prediktivni modeli. Za regresijski model odabrana je varijabla PageValues, jer je riječ o kontinuiranoj mjeri koja opisuje procijenjenu vrijednost posjeta (što je posjet bliži kupnji, to je vrijednost veća). Takva varijabla je prikladna za regresiju jer model može učiti koliko se ta vrijednost mijenja ovisno o ponašanju korisnika na stranici (broj pregledanih proizvoda, trajanje posjeta, tip posjetitelja i sl.). Za klasifikacijski model odabrana je varijabla Revenue, koja već postoji u skupu i ima samo dvije vrijednosti (kupnja je ostvarena/kupnja nije ostvarena), pa se može izravno koristiti za zadatak klasifikacije.

5. Prediktivno modeliranje

5.1. Priprema podataka

U fazi prediktivnog modeliranja skup podataka je pripremljen tako da se odvoje ulazne značajke (X) i izlazna varijabla (y). Budući da projekt uključuje i klasifikaciju i regresiju, napravljene su dvije zasebne podjele, pri čemu se u svakoj podjeli kao ciljna varijabla uzima drugačiji stupac, a sve ostalo ulazi u X. Detaljna objašnjenja značenja svih varijabli već su dana na početku projekta (u poglavlju gdje su opisane varijable skupa podataka), pa se ovdje navodi samo koje se varijable koriste u kojem modelu.

5.1.1. Podjela podataka na ulazne i izlazne varijable

Klasifikacijski model (cilj: Revenue)

Za klasifikaciju je izlazna varijabla Revenue (0 = nema kupnje, 1 = kupnja), pa je y = Revenue. Ulazne značajke su sve preostale varijable osim ciljne i osim PageValues. To konkretno uključuje: Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration, BounceRates, ExitRates, SpecialDay, Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend.

Varijabla PageValues namjerno nije uključena u klasifikaciju jer je snažno povezana s ishodom kupnje i njezino uključivanje bi moglo uzrokovati “curenje informacija”, odnosno nerealno dobru uspješnost modela.

Regresijski model (cilj: PageValues)

Za regresiju je izlazna varijabla PageValues, pa je y = PageValues. Ulazne značajke su sve ostale varijable osim ciljne i osim Revenue. To konkretno uključuje iste varijable kao što su navedene u tekstu iznad.

Varijabla Revenue isključena je iz regresije jer predstavlja konačni ishod kupnje i zbog jake povezanosti s PageValues, može umjetno povećati uspješnost modela.

5.1.2. Podjela na trening i testni skup

Podaci su podijeljeni na trening i testni skup kako bismo mogli objektivno provjeriti koliko dobro modeli rade na novim, neviđenim podacima. Korišten je omjer 80% / 20% (80% podataka za učenje, 20% za evaluaciju). Podjela je provedena funkcijom train_test_split uz postavljeni random_state=42 kako bi rezultati bili ponovljivi.

5.2. Treniranje i evaluacija - klasifikacija (Revenue)

Za klasifikacijski zadatak (predviđanje varijable Revenue) korištena je logistička regresija. Model ne daje odmah “0 ili 1”, nego prvo računa vjerojatnost kupnje, a zatim na temelju praga odlučuje hoće li ishod biti kupnja (1) ili ne (0).

Prije treniranja, kategorijske varijable Month, VisitorType, Weekend, OperatingSystems, Browser, Region i TrafficType pretvorene su u numerički oblik pomoću one-hot encodinga, dok su numeričke značajke skalirane metodom StandardScaler kako bi model stabilnije učio i kako pojedine varijable ne bi dominirale samo zato što su na većoj skali. Prilikom podjele podataka korišten je i stratify=y kako bi omjer klasa (kupnja/nema kupnje) ostao približno jednak u trening i testnom skupu, što je važno jer je klasa kupnje rjeđa (oko 15-16% slučajeva).

Kod evaluacije nije dovoljno gledati samo accuracy, jer kod neuravnoteženih podataka model može imati visoku točnost i ako uglavnom predviđa većinsku klasu

(nema kupnje). Zbog toga su korištene i metrike koje bolje opisuju uspješnost prepoznavanja kupnje: precision (koliko su predviđene kupnje stvarno točne), recall (koliki dio svih stvarnih kupnji je model prepoznao) i F1-score (kompromis precision i recall) za klasu 1, te ROC AUC i PR AUC kao mjere razdvajanja klasa i kvalitete rangiranja vjerojatnosti, pri čemu je PR AUC posebno informativan kada je pozitivna klasa rijetka.

U projektu su napravljene dvije verzije logističke regresije. Basic LR je osnovni model s početnim postavkama i zadanim pragom 0,5. Na ovom skupu podataka postiže visoku accuracy jer uglavnom dobro predviđa većinsku klasu (nema kupnje), ali pritom propušta velik broj stvarnih kupnji pa ima niži recall za klasu 1.

Kako bi se dobila bolja verzija modela, izrađen je Best LR pomoću postupka GridSearchCV. GridSearchCV automatski isprobava više kombinacija hiperparametara logističke regresije i za svaku kombinaciju radi cross-validation unutar trening skupa, a zatim odabire onu kombinaciju koja daje najbolji rezultat prema unaprijed odabranoj metrici. U ovom slučaju koristili smo F1-score jer najbolje pokazuje koliko dobro model prepoznaje kupnju, koja je rijetka u podacima. GridSearchCV je kao najbolje postavke odabrao penalty='l2', C=1 i class_weight='balanced' pa je Best LR definiran kao logistička regresija s tim postavkama. Parametar class_weight='balanced' povećava važnost rijetke klase (kupnje), pa je u ovom projektu Best LR ostvario veći recall i F1-score za klasu 1, ali uz cijenu niže ukupne accuracy. U kontekstu web-trgovine takav kompromis često ima smisla, jer je korisnije prepoznati više stvarnih kupnji, čak i ako ukupna točnost padne, nego imati visoku točnost samo zato što model većinu vremena predviđa “nema kupnje”, jer je to najčešći slučaj u podacima.

5.3 Treniranje i evaluacija - regresija (PageValues)

Za regresijski zadatak kao ciljna varijabla koristi se PageValues, koja je kontinuirana (raspon 0-361.76), ali je pritom vrlo neuravnotežena jer 77.63% vrijednosti iznosi 0. To predstavlja izazov za standardne regresijske modele (tzv. zero-inflation), jer velik udio nultih vrijednosti može navesti model da predviđa vrijednosti bliže nuli.

U pripremi podataka, kategorijske varijable Month, VisitorType, Weekend, OperatingSystems, Browser, Region i TrafficType kodirane su metodom one-hot encoding (uz drop_first=True, kako bi se izbjegla suvišna/dupla kodiranja). Nakon pripreme, modeli su trenirani na trening skupu te evaluirani na testnom skupu pomoću regresijskih metrika koje su prikladne za procjenu pogreške predikcije.

5.3.3 Odabir algoritama

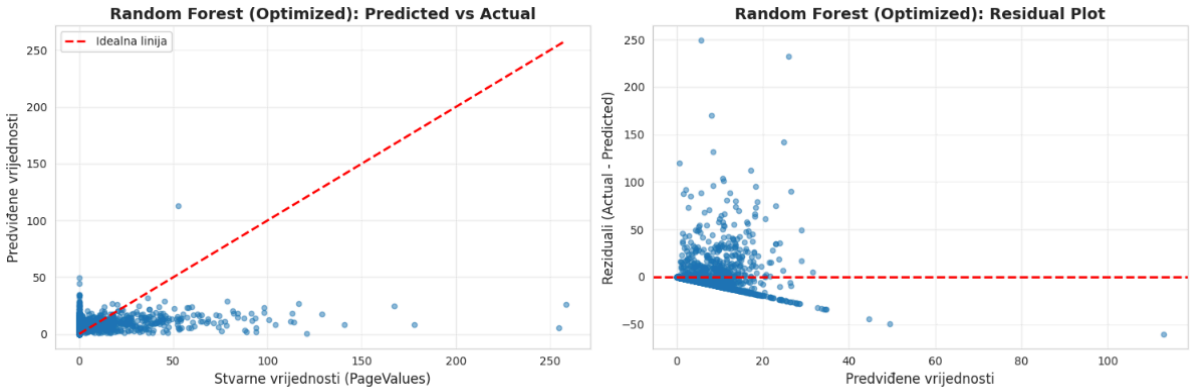
Za regresijsko modeliranje odabrana su tri tree-based algoritma zbog njihove robusnosti na outliere i sposobnosti hvatanja nelinearnih odnosa:

1. Random Forest Regressor - ensemble metoda koja kombinira višestruka stabla odlučivanja
2. XGBoost Regressor - gradient boosting algoritam s ugrađenom regularizacijom
3. Gradient Boosting Regressor - sekvencijalno učenje kroz minimizaciju gubitka

5.3.4 Treniranje modela

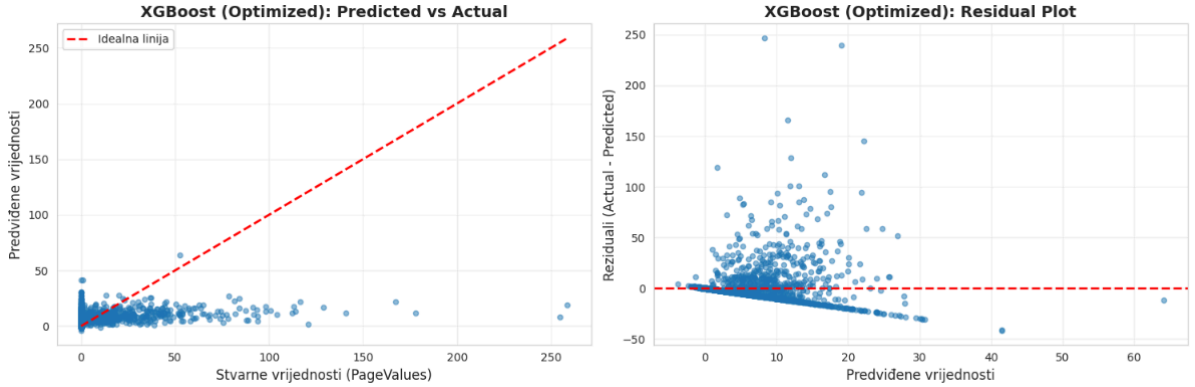
Svaki model je prvo treniran s osnovnim parametrima, zatim je provedena optimizacija hiperparametara korištenjem GridSearchCV s 5-fold cross-validation. Evaluacija je izvršena na temelju negative mean squared error kao primarnog scoring kriterija.

Random Forest optimalni parametri: max_depth=10, min_samples_leaf=2, n_estimators=200



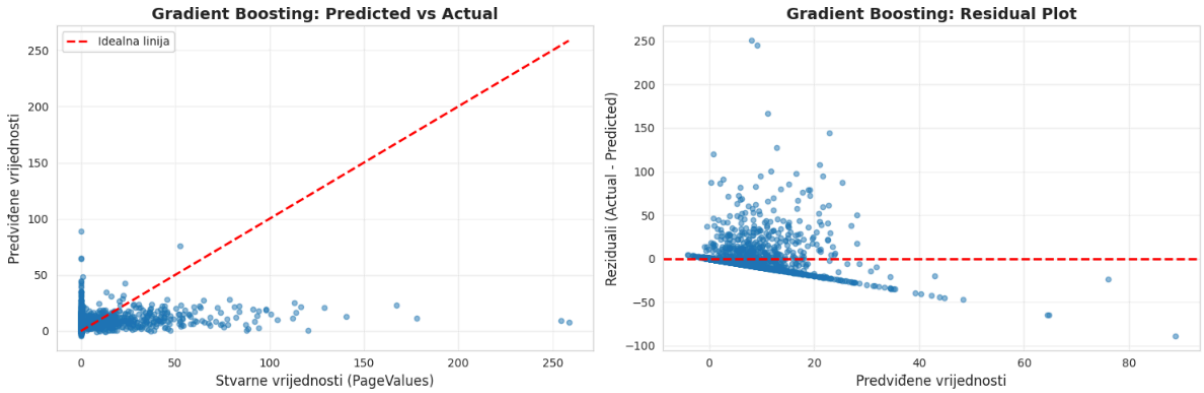
Slika 29: Random Forest – usporedba predviđenih i stvarnih vrijednosti

XGBoost optimalni parametri: max_depth=3, learning_rate=0.1, n_estimators=100, subsample=1.0



Slika 30: XGBoost – usporedba predviđenih i stvarnih vrijednosti

Gradient Boosting optimalni parametri: max_depth=3, learning_rate=0.05, n_estimators=200, subsample=0.8



Slika 31:Gradient Boosting – usporedba predviđenih i stvarnih vrijednosti

5.3.5 Evaluacija modela

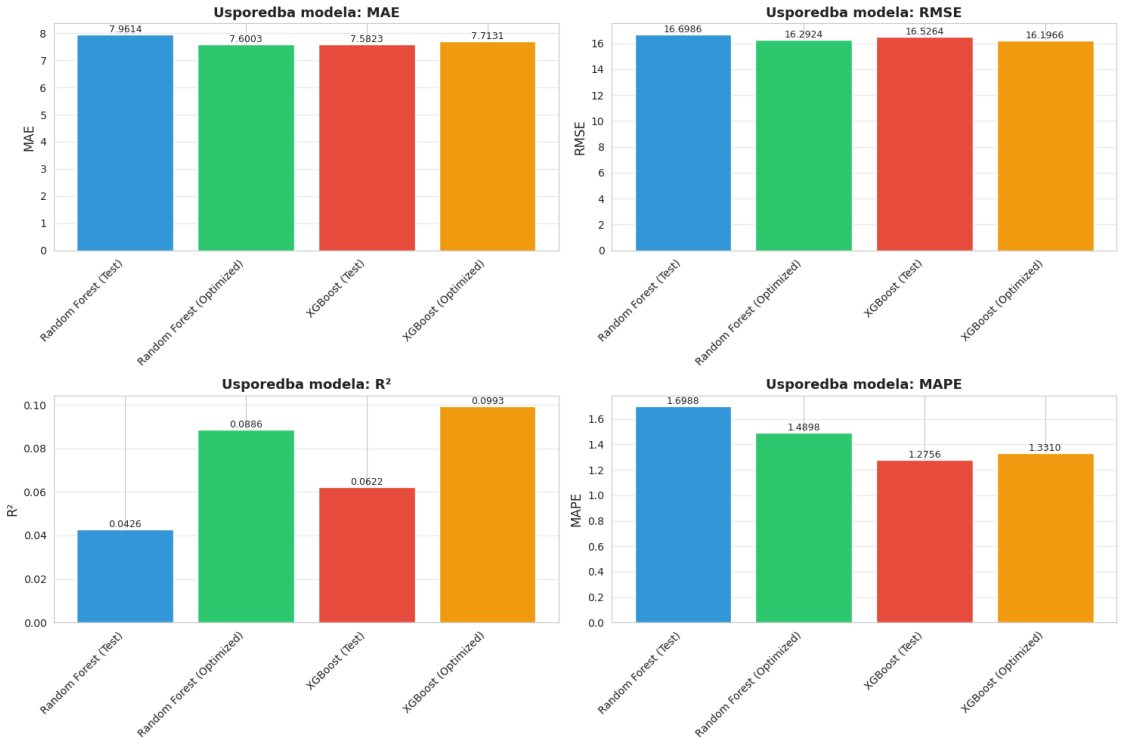
Modeli su evaluirani korištenjem standardnih regresijskih metrika:

Model	R^2	RMSE	MAE	MAPE
Random Forest (Opt.)	0.0886	16.29	7.60	1.49
XGBoost (Opt.)	**0.0993**	**16.20**	7.71	1.33
Gradient Boosting (Opt.)	0.0883	16.30	7.73	1.34

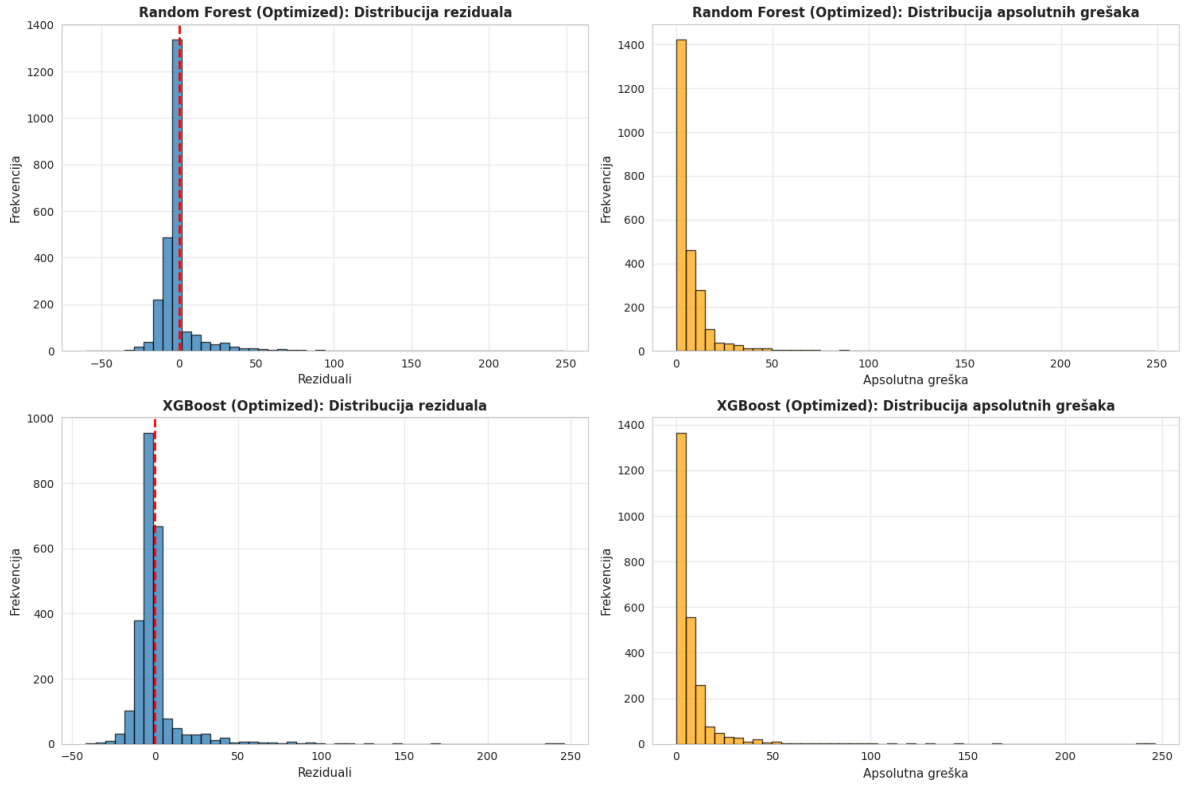
Tablica 1. Evaluacija regresijskih modela

XGBoost (Optimized) postiže najbolje rezultate s $R^2 = 0.0993$, što znači da model objašnjava približno 10% varijance ciljne varijable. RMSE od 16.20 ukazuje na prosječnu kvadratnu grešku predviđanja, što je značajno s obzirom da je srednja vrijednost PageValues 5.95.

Sva tri modela pokazuju konzistentne rezultate (razlika u R^2 manja od 0.02), što ukazuje da niska prediktivna moć nije posljedica odabira algoritma, već prirode podataka.



Slika 32: Usporedba performansi svih modela prema ključnim metrikama



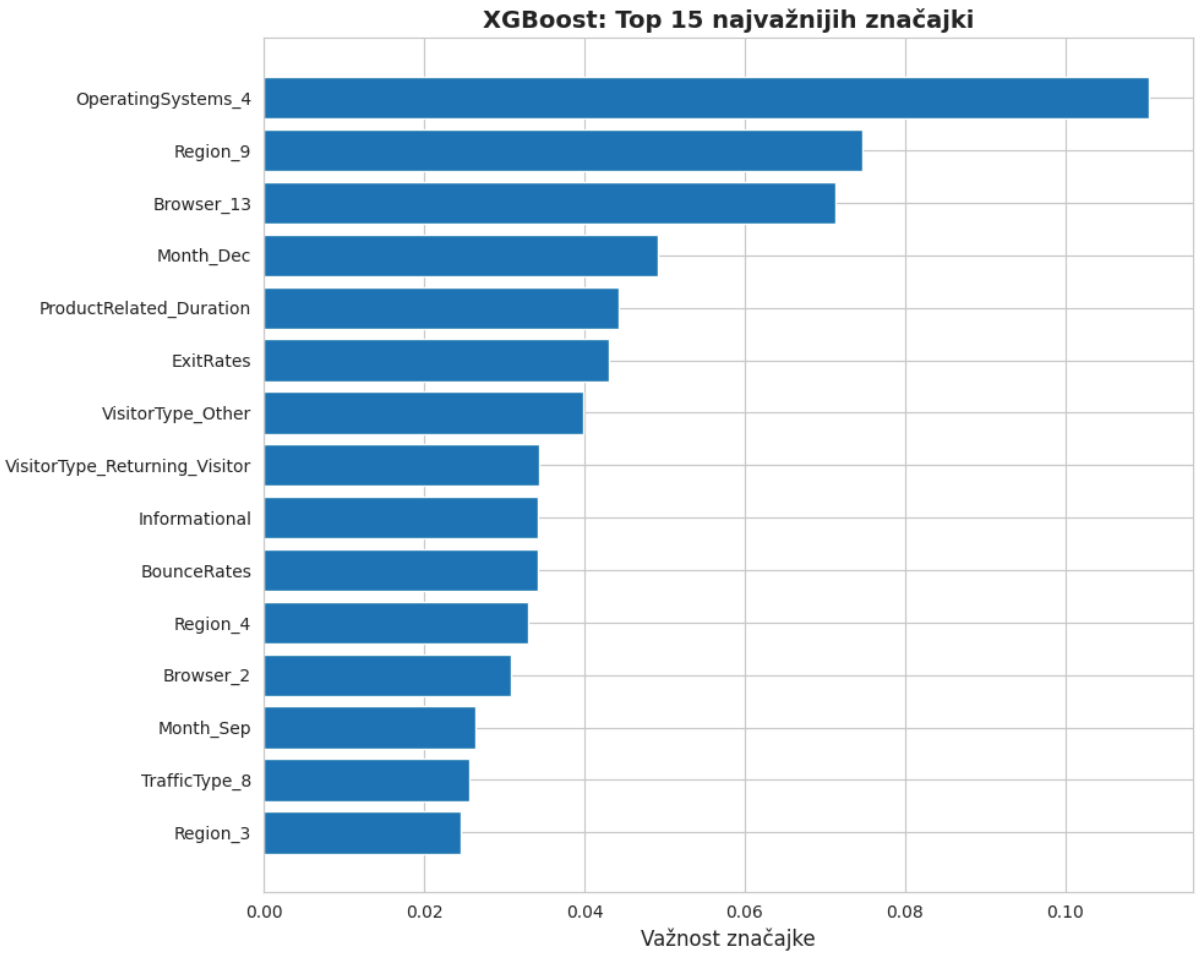
Slika 33: Analiza distribucije grešaka optimiziranih modela

5.3.6 Analiza značajki

Analiza važnosti značajki pokazuje različite prioritete ovisno o algoritmu:

XGBoost model identificira kategorijske značajke kao dominantne:

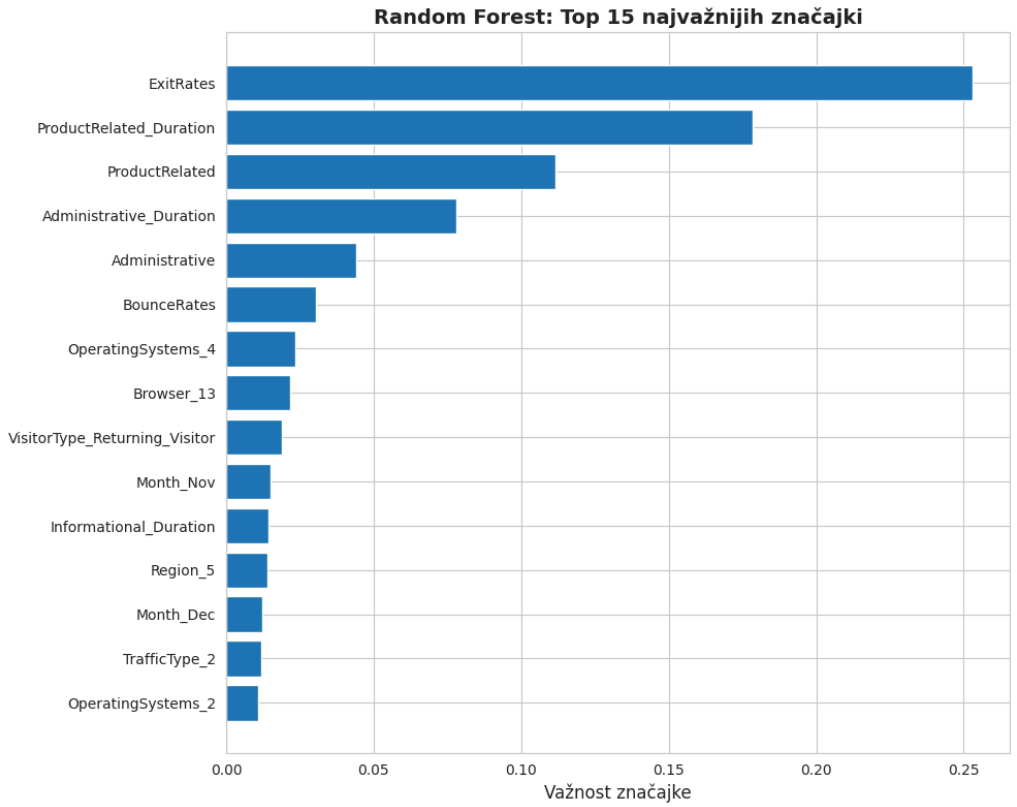
- OperatingSystems_4 (0.112) - najvažnija značajka
- Region_9 (0.075) - geografska regija
- Browser_13 (0.068) - tip web preglednika
- Month_Dec (0.050) - prosinci (sezonalnost)
- ProductRelated_Duration (0.043) - ponašanje na stranicama proizvoda



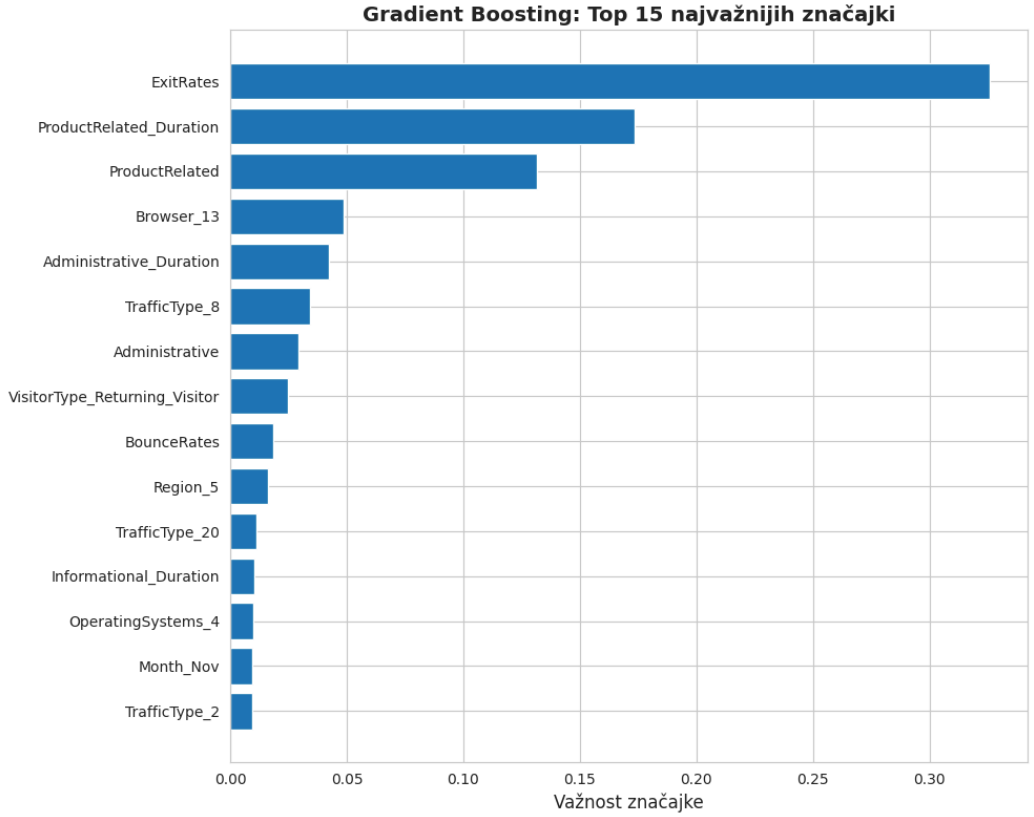
Slika 34: Top 15 najvažnijih značajki prema XGBoost modelu.

Random Forest i Gradient Boosting daju prioritet ponašajnim metrikama:

- ExitRates - dominantno najvažnija značajka
- ProductRelated_Duration - vrijeme na stranicama proizvoda
- ProductRelated - broj pregledanih proizvoda
- Administrative_Duration, BounceRates - ostale ponašajne metrike



Slika 35: Top 15 najvažnijih značajki prema Random Forest modela



Slika 36: Top 15 najvažnijih značajki prema Gradient Boosting modelu

Interpretacija

Razlika u prioritetima ukazuje na različite pristupe. XGBoost hvata segmente korisnika (OS/Browser/Region kombinacije). Tree-based ansambli (RF, GB) fokusiraju se na stvarno ponašanje na primjer koliko dugo ostaju, koliko brzo napuštaju

Oba pristupa su komplementarna i zajedno daju potpuniju sliku segmentacije prema tehno-profilu (XGBoost) i optimizacija ponašanja na stranici (RF/GB)

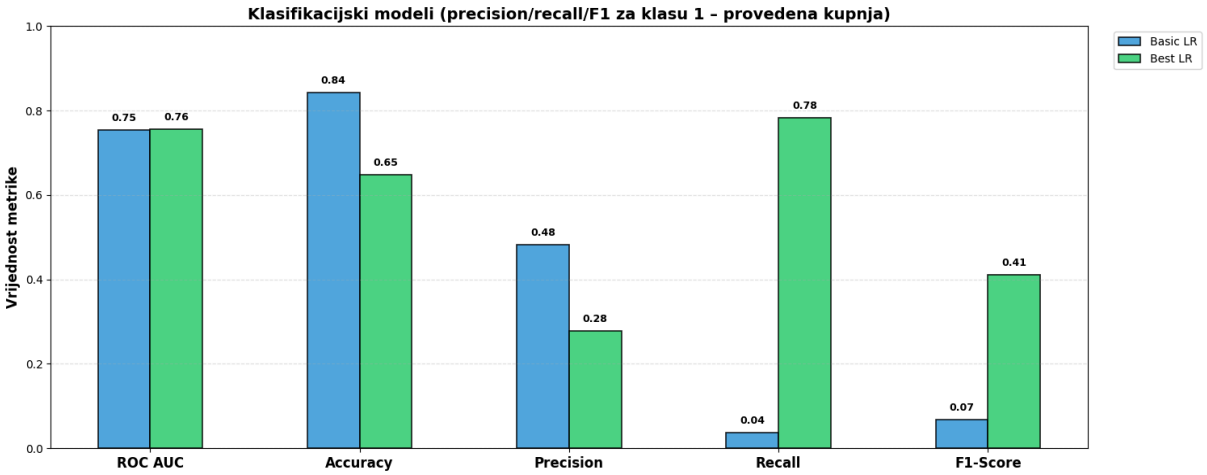
5.4. Evaluacija i usporedba modela

Usporedba klasifikacijskih modela provedena je na testnom skupu pomoću metrika Accuracy, Precision, Recall i F1-score za klasu 1 (kupnja), te ROC AUC i PR AUC kao mjera kvalitete rangiranja vjerojatnosti.

Basic LR na testu postiže Accuracy = 0.8431 i ROC AUC = 0.7544 (PR AUC = 0.3329), ali pri zadanom pragu 0.5 model rijetko predviđa kupnju. To se vidi iz matrice zabune (TP=14, FP=15, FN=368), zbog čega su metrike za klasu kupnje niske: Recall(1)=0.0366 i F1(1)=0.0681, iako je Precision(1)=0.4828.

Best LR je dobiven optimizacijom hiperparametara pomoću GridSearchCV, pri čemu su odabrane postavke penalty='l2', C=1, class_weight='balanced' jer su dale najbolji rezultat prema F1-scoreu u unakrsnoj provjeri na trening skupu. Na testu Best LR postiže Accuracy = 0.6489, ROC AUC = 0.7557 (PR AUC = 0.3312), ali značajno bolje prepoznaje kupnje: iz matrice zabune (TP=299, FP=774, FN=83) dobivamo Recall(1)=0.7827, Precision(1)=0.2787 i F1(1)=0.4110.

Budući da je cilj zadatka prepoznavanje kupnji (rijetke klase), Best LR je prikladniji izbor: iako ima nižu ukupnu točnost, mnogo bolje pronalazi korisnike koji će kupiti, što se jasno vidi kroz znatno veći recall i F1-score za klasu 1.

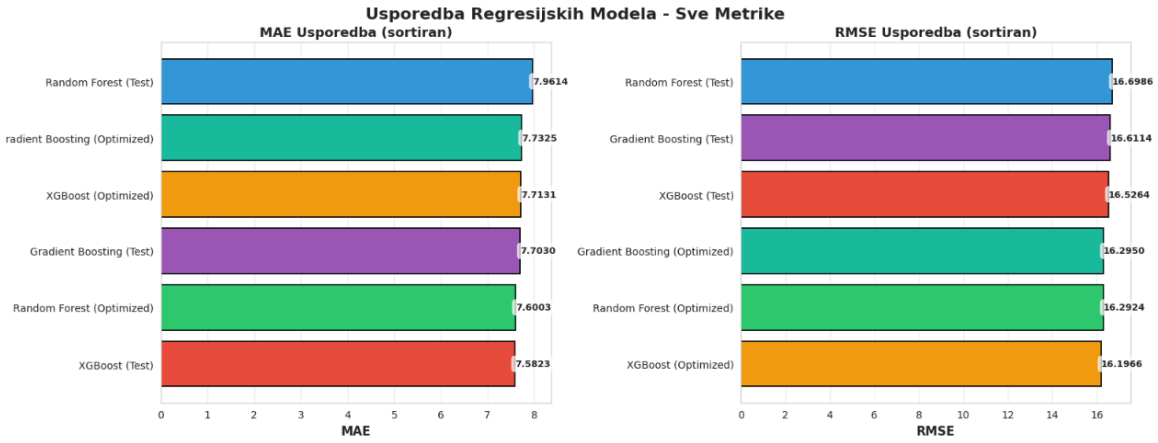


Slika 37: Metrike klasifikacijskih modela

Usporedba regresijskih modela provedena je na testnom skupu pomoću metrika MAE i RMSE, koje mjere prosječnu pogrešku predviđanja na originalnoj skali varijable *PageValues*. Na slici 38. prikazane su pogreške za optimizirane modele Random Forest, XGBoost i Gradient Boosting.

Random Forest (Optimized) ostvaruje RMSE = 16.29 i MAE = 7.60, što znači da model u prosjeku griješi oko 9 jedinica (MAE), dok RMSE pokazuje nešto veću ukupnu pogrešku jer jače "kažnjava" rijetke slučajeve s velikim odstupanjima. XGBoost (Optimized) postiže najniži RMSE = 16.20, uz MAE = 7.71 i najviši R^2 0.0993, pa je najbolji prema kriteriju koji više naglašava veće pogreške. Gradient Boosting (Optimized) daje RMSE = 16.30 i MAE = *7.73*, vrlo blizu ostalim modelima.

Sva tri optimizirana modela postižu vrlo slične R^2 vrijednosti (0.0883-0.0993), što znači da objašnjavaju približno 9-10% varijance ciljne varijable.



Slika 38: Regresijski modeli (RMSE/MAE)

Razlike među modelima su minimalne (RMSE se razlikuje za svega ~0.04, a MAE za ~0.06), što upućuje na to da izbor konkretnog tree-based algoritma ovdje nema veliki utjecaj na konačnu točnost. Takav rezultat je očekivan jer je *PageValues* izrazito "teška" ciljna varijabla (velik udio nultih vrijednosti i mali broj vrlo velikih vrijednosti), pa modeli teško precizno pogađaju rijetke visoke iznose, što se posebno vidi kroz relativno visok RMSE. Ako se kao prioritet uzme stabilnija prosječna pogreška, tada je Random Forest blago najbolji po MAE, dok je XGBoost najbolji po RMSE (najmanje velikih promašaja). U praksi se može odabrati XGBoost kao najuravnoteženiji izbor prema RMSE, uz napomenu da bi se značajnije poboljšanje najvjerojatnije postiglo promjenom pristupa (npr. dvofazno modeliranje: prvo predvidjeti je li PageValues = 0 ili > 0, a zatim regresijom procjenjivati vrijednost samo za slučajeve > 0).

6. Interpretacija rezultata i zaključci

6.1. Usporedba rezultata s drugim radovima

U ranijim radovima cilj je isti: iz ponašanja korisnika na web-trgovini procijeniti hoće li doći do kupnje. Jain (2025) pokazuje da modeli koji prate redoslijed klikova kroz vrijeme mogu bolje uhvatiti obrasce ponašanja, a Eudoxus Press (2024) na istom UCI skupu pokazuje da napredniji modeli često daju bolje rezultate od jednostavnih, ali upozorava da se ponekad rezultati mogu činiti boljima nego što stvarno jesu ako se model previše prilagodi podacima. Druga dva rada na anketama naglašavaju da kupnja ne ovisi samo o klikovima nego i o iskustvu, navikama i kontekstu, te da se odnos između "namjere" i stvarne kupnje može razlikovati među skupinama ljudi.

U usporedbi s time, dobiveni rezultati klasifikacije pokazuju da optimizirana logistička regresija daje korisniji kompromis za prepoznavanje kupnje: u odnosu na osnovni model prepoznaje znatno više stvarnih kupnji (veći recall i F1-score za klasu 1), ali pritom povećava broj lažnih pozitivnih predikcija, odnosno situacija kada model predvidi kupnju iako je nije bilo. Unatoč tome, realno je očekivati da bi napredniji pristupi iz literature mogli dodatno poboljšati prepoznavanje kupaca. Kod predviđanja PageValues rezultati su slabiji i to je očekivano, jer većina sesija ima PageValues = 0

pa je vrijednost teško precizno pogoditi; zato bi se to moglo poboljšati tako da se prvo procijeni hoće li vrijednost biti 0 ili ne, a tek onda predviđa kolika je vrijednost za one sesije gdje nije nula.

6.2. Zaključak i smjernice za dalje

Kroz projekt se pokazalo da su podaci uredni i pouzdani za rad (nema nedostajućih vrijednosti, uklonjeni su duplikati), ali da priroda problema nameće dva glavna ograničenja: varijabla Revenue je neizbalansirana, a PageValues ima vrlo velik udio nultih vrijednosti. Zbog toga je klasifikacijski dio dao upotrebljive rezultate, dok je regresijski dio očekivano teži i slabije predvidljiv.

Za daljnje unaprjeđenje analize, u klasifikaciji bi imalo smisla proširiti ispitivanje na naprednije modele koji su se u literaturi pokazali učinkovitima na sličnim podacima, jer je realno očekivati poboljšanje u odnosu na logističku regresiju. Zbog neuravnoteženosti podataka naglasak evaluacije treba ostati na metrikama poput F1-score i PR AUC koje bolje opisuju uspješnost prepoznavanja kupnje nego sama točnost.

Kod regresije PageValues glavni problem nije u izboru algoritma nego u tome što većina vrijednosti iznosi 0, pa standardni regresijski modeli teško uče rijetke veće vrijednosti. Zato bi sljedeći korak trebao biti promjena pristupa: prvo procijeniti hoće li PageValues biti 0 ili veći od 0, a zatim predviđati samu vrijednost samo za one sesije gdje je PageValues različit od nule. Takav pristup bi bolje odgovarao stvarnoj strukturi ciljne varijable i potencijalno dao stabilnije rezultate.

Dodatno, korisno bi bilo povezati rezultate klasteriranja s prediktivnim modelima, odnosno provjeriti u kojim se skupinama korisnika model najviše griješi i što te skupine karakterizira. Time bi se dobile jasnije i praktičnije preporuke za optimizaciju sadržaja web-trgovine i izvora prometa. Ako bi u budućnosti bili dostupni detaljniji podaci o slijedu događaja unutar sesije (redoslijed pregleda, dodavanje u košaricu, povratak na stranice i slično), njihovo uključivanje bi moglo dodatno poboljšati rezultate, jer ranija istraživanja pokazuju da takve vremenske informacije često nose snažan signal za predviđanje kupnje.

Popis literature

[1] A. Jain, “Predicting E-commerce Purchase Behavior using a DQN-Inspired Deep Learning Model for enhanced adaptability,” *arXiv preprint*, arXiv:2506.17543, 2025. [Online]. Available: https://arxiv.org/pdf/2506.17543

[2] *Understanding Online Shoppers’ Purchase Intentions using Data Analytics*, Eudoxus Press, 2024. [Online]. Available: https://eudoxuspress.com/index.php/pub/article/view/1892/1218

[3] J. Kim, H. Lee, and H. Kim, “Factors Affecting Online Search Intention and Online Purchase Intention,” *Seoul National University*, Dec. 2004. [Online]. Available: https://s-space.snu.ac.kr/handle/10371/1809

[4] N. Peña-García, I. Gil-Saura, A. Rodríguez-Orejuela, and J. R. Siqueira-Junior, “Purchase intention and purchase behavior online: A cross-cultural approach,” *Heliyon*, vol. 6, e04284, 2020. [Online]. Available: https://www.cell.com/heliyon/fulltext/S2405-8440(20)31128-2

Popis slika