https://doi.org/10.37528/FTTE/9788673954165/POSTEL.2022.026

PREDIKCIJA ODLIVA KORISNIKA TELEKOMUNIKACIONIH OPERATORA PRIMENOM MAŠINSKOG UČENJA

Slađana Janković[1], Snežana Mladenović[2], Ivana Stefanović[3], Ana Uzelac[4]
[1]Univerzitet u Beogradu - Saobraćajni fakultet, s.jankovic@sf.bg.ac.rs
[2]Univerzitet u Beogradu - Saobraćajni fakultet, snezanam@sf.bg.ac.rs
[3]Akademija tehničko-umetničkih strukovnih studija Beograd – Odsek Visoka škola elektrotehnike i računarstva, ivanas@viser.edu.rs
[4]Univerzitet u Beogradu - Saobraćajni fakultet, ana.uzelac@sf.bg.ac.rs

Rezime: *Kako bi prevenirao odliv korisnika, za telekomunikacionog operatora bilo bi korisno da sazna koji su to parametri koji najviše utiču na odlazak korisnika. Rad se bavi problemom predikcije budućeg odliva korisnika na osnovu istorijskih podataka u programskom jeziku Python. U cilju rešavanja ovog problema pronađen je odgovarajući, otvoreni skup podataka i izvršena istraživačka analiza podataka, kako bi se utvrdio stepen zavisnosti između svake nezavisne i zavisne varijable. Nezavisne varijable opisuju korisnika i servise koje je koristio, dok zavisna varijabla daje odgovor na pitanje: da li je korisnik do tog trenutka napustio operatora? Zatim su kreirani različiti klasifikacioni modeli mašinskog učenja korišćenjem nekih od algoritama implementiranih u Scikit-Learn biblioteci programskog jezika Python. Tačnost najboljih modela iznosila je preko 95%, što je za 10% više od tačnosti null modela, pa se može zaključiti da se predikacija odliva korisnika može uspešno vršiti korišćenjem mašinskog učenja, u programskom jeziku Python.*

Ključne reči: *mašinsko učenje, odliv korisnika, predikcija, klasifikacija, Python*

1. Uvod

Telekomunikacioni operatori ulažu velike napore kako bi privukli što veći broj novih korisnika i kako bi što duže zadržali postojeće korisnike na visoko konkurentnom tržištu. Trošak zadržavanja postojećih korisnika je obično niži u poređenju sa troškom privlačenja novih korisnika [1]. Upravo iz ovog razloga, razvijen je veliki broj različitih modela koji omogućavaju predikciju budućeg odliva korisnika u cilju preveniranja istog. U ovom istraživanju, na osnovu istorijskih podataka o odlivu korisnika, u programskom jeziku *Python*, obučeni su i verifikovani modeli mašinskog učenja, uz pomoć kojih se sa visokom tačnošću može predvideti za koje korisnike se očekuje da će napustiti telekomunikacionog operatora.

U drugoj sekciji rada opisana je metodologija ovog istraživanja, tj. sve faze procesa mašinskog učenja, realizovanog u cilju predikcije odliva korisnika telekomunikacionih operatora. U trećoj sekciji rada prezentovani su rezultati prediktivne

analize sprovedene na raspoloživom skupu podataka i dobijeni rezultati su upoređeni sa rezultatima sličnih istraživanja. Četvrta sekcija sadrži najznačajnije zaključke koji su proistekli iz ovog istraživanja.

2. Metodologija

Ovo istraživanje obuhvata primenu metode nadgledanog mašinskog učenja, u rešavanju problema binarne klasifikacije korisnika telekomunikacionog operatora. Korisnici se klasifikuju prema tome da li su do tog trenutka napustili posmatranog operatora ili nisu. Istraživačka analiza i vizuelizacija podataka, priprema skupa podataka za proces mašinskog učenja, kao i izgradnja i primena modela mašinskog učenja urađeni su u programskom jeziku *Python*, na *Jupyter Notebook web* platformi.

2.1. Skup podataka

U istraživanju je korišćen otvoreni skup podataka kompanije *Orange* [1]. Skup podataka opisuje profile korisnika nepoznatog operatora mobilne telefonije iz Sjedinjenih Američkih Država (SAD). Skup podataka sastoji se od 3333 zapisa, pri čemu svaki zapis opisuje jednog korisnika. Ukupan broj atributa u skupu podataka je 21. Originalni skup podataka je na engleskom jeziku [2], ali je za potrebe ovog istraživanja preveden na srpski jezik. Prevođenjem tekstualnih vrednosti na srpski jezik nisu promenjeni originalni tipovi podataka atributa. U Tabelama 1, 2, 3 i 4 prikazano je prvih pet instanci skupa podataka koji je korišćen za predikciju u ovom istraživanju. Poslednja kolona – odliv korisnika, je ciljna varijabla, dok su ostale kolone nezavisni atributi. Ciljna varijabla ima vrednost *TRUE* ukoliko je posmatrani korisnik otkazao pretplatu, tj. napustio ovog operatora, odnosno vrednost *FALSE*, ukoliko nije. Samo vrednosti ciljne varijable nisu prevedene na srpski jezik, kako bi i ova varijabla ostala istog tipa podataka kao u originalnom skupu podataka, tj. logičkog tipa (tip *bool*). Skup podataka dobijen prevođenjem na srpski jezik smatraće se u nastavku istraživanja izvornim skupom podataka.

Tabela 1. Prvih pet instanci izvornog skupa podataka – prvi deo

država u SAD	**trajanje računa u danima**	**pozivni broj**	**broj telefona**	**plan međunarodnih poziva**
KS	128	415	382-4657	ne
OH	107	415	371-7191	ne
NJ	137	415	358-1921	ne
OH	84	408	375-9999	da
OK	75	415	330-6626	da

Tabela 2. Prvih pet instanci izvornog skupa podataka – drugi deo

mogućnost glasovne pošte	**broj glasovnih poruka**	**ukupan broj minuta dnevno**	**ukupan broj poziva dnevno**	**ukupna cena dnevnih poziva**
da	25	265.1	110	45.07
da	26	161.6	123	27.47
ne	0	243.4	114	41.38
ne	0	299.4	71	50.90
ne	0	166.7	113	28.34

Tabela 3. Prvih pet instanci izvornog skupa podataka – treći deo

ukupan broj večernjih minuta	ukupan broj večernjih poziva	ukupna cena večernjih poziva	ukupan broj noćnih minuta	ukupan broj noćnih poziva
197.4	99	16.78	244.7	91
195.5	103	16.62	254.4	103
121.2	110	10.30	162.6	104
61.9	88	5.26	196.9	89
148.3	122	12.61	186.9	121

Tabela 4. Prvih pet instanci izvornog skupa podataka – četvrti deo

ukupna cena noćnih poziva	ukupan broj međunaro-dnih minuta	ukupan broj međunaro-dnih poziva	ukupna cena međuna-rodnih poziva	broj poziva korisni-čkog servisa	odliv korisnika
11.01	10.0	3	2.70	1	FALSE
11.45	13.7	3	3.70	1	FALSE
7.32	12.2	5	3.29	0	FALSE
8.86	6.6	7	1.78	2	FALSE
8.41	10.1	3	2.73	3	FALSE

2.2. Istraživačka analiza i priprema podataka

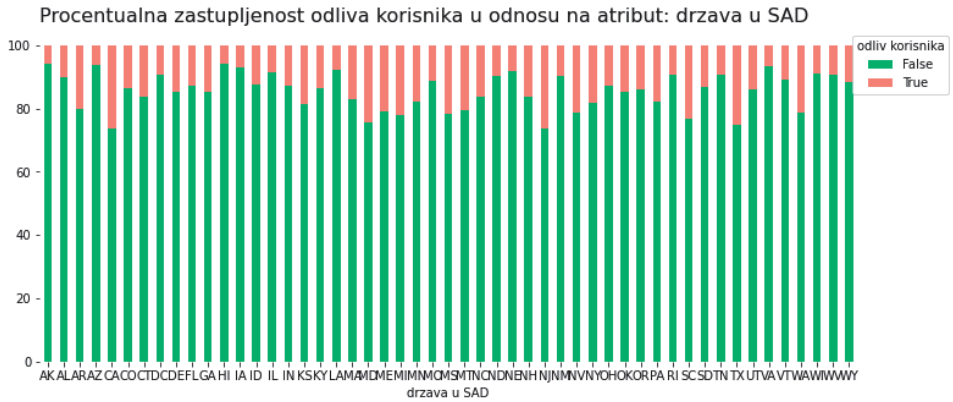
Istraživačka analiza podataka (engl. *Exploratory Data Analysis*) radi se iz više razloga: da bi se utvrdile osnovne osobine atributa, kao što su tipovi podataka, nedostajuće vrednosti i jedinstvene vrednosti, da bi se utvrdio značaj svakog nezavisnog atributa, tj. stepen u kojem vrednost ciljne varijable zavisi od njegove vrednosti i na kraju, da bi se izvorni skup podataka na pravi način pripremio za proces mašinskog učenja.

Skup podataka skladišten je u *.csv* datoteci *odliv korisnika u telekomunikacijama.csv*, koja je postavljena na *Jupyter Notebook* platformu. Podaci iz ove datoteke pročitani su i smešteni u *pandas* okvir podataka (engl. *Data Frame*) *df_telekomunikacije_odliv_korisnika*. *pandas* je softverska biblioteka pisana za programski jezik *Python*, namenjena za manipulisanje i analizu podataka u formi sekvenca podataka i u formi tabela podataka. Okvir podataka predstavlja dvodimenzionalnu (tabelarnu) strukturu podataka. Upoznavanje sa izvornim skupom podataka je veoma važno u primeni metode mašinskog učenja, jer je preduslov za odgovarajuću pripremu raspoloživog skupa podataka za kreiranje modela mašinskog učenja. Naredbom *df_telekomunikacije_odliv_korisnika.info()* dobijen je izveštaj o osnovnim karakteristikama skupa podataka, koji će u nastavku biti korišćen za predikciju. Iz izveštaja se moglo videti da u izvornom skupu podataka nije bilo *null* vrednosti, pa nije bilo potrebe za procenom i imputiranjem nedostajućih vrednosti, isključivanjem atributa koji ih sadrže ili eliminisanjem nepotpunih instanci. Za pripremu skupa podataka za prediktivnu analizu, kao i za vizuelizaciju podataka, veoma je važan uvid u tipove podataka nezavisnih promenljivih, kao i ciljne (zavisne) promenljive. U pomenutom

izveštaju mogu se videti i *pandas* tipovi podataka promenljivih: jedna promenljiva, i to ciljna, je tipa *bool* (logička), osam promenljivih su tipa *float64* (numeričke realne), osam promenljivih su tipa *int64* (numeričke celobrojne) i četiri promenljive su tipa *object* (mogu biti tekstualne ili mešovite - tekstualne i numeričke).

Naredbom *df_telekomunikacije_odliv_korisnika.nunique()* određen je i prikazan broj jedinstvenih vrednosti u svakoj koloni skupa podataka, odnosno okvira podataka. Za atribut *broj telefona* prikazano je da ima 3333 jedinstvenih vrednosti, tj. onoliko vrednosti koliko ukupno ima zapisa. To znači da ovaj atribut nema nikakav uticaj na vrednost ciljne promenljive, pa je iz tog razloga on i obrisan iz okvira podataka, sledećom naredbom: *df_telekomunikacije_odliv_korisnika.drop(columns=['broj telefona'], inplace=True)*.

Korišćenjem odgovarajućih funkcija i metoda *Python* biblioteka i modula, kao što su: *numpy*, *math*, *matplotlib* i *plotly*, iscrtani su dijagrami prikazani na Slikama 1 i 2. Na Slici 1 prikazana je procentualna raspodela odliva korisnika za svaku državu SAD. Ovaj grafikon pokazuje da odliv korisnika po državama SAD varira približno između 5% i 25%, što ukazuje na to da *drzava u SAD*, kao atribut, ne utiče značajno na vrednost ciljne varijable. Iz tog razloga je i ovaj atribut isključen, tj. obrisan iz okvira podataka.

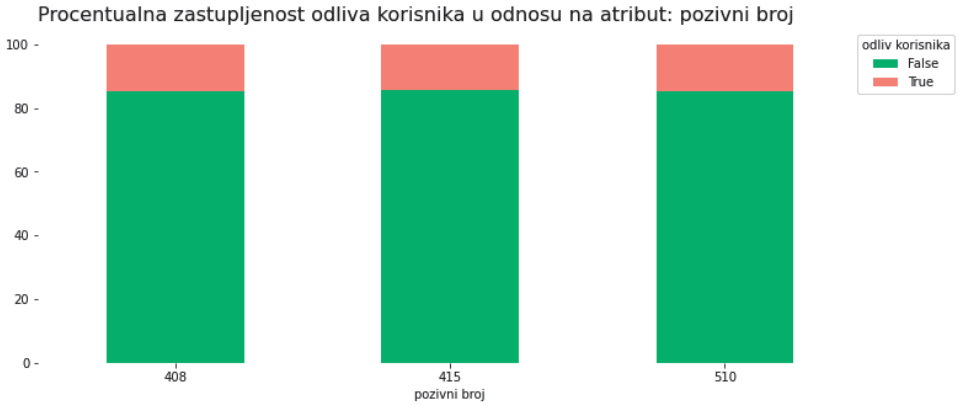


Slika 1. Procentualna raspodela odliva korisnika po državama SAD

Na Slici 2 prikazana je procentualna raspodela odliva korisnika za svaki od tri pozivna broja, koji su pronađeni u skupu podataka. Ovaj grafikon pokazuje da je odliv korisnika po pozivnim brojevima veoma ujednačen i iznosi približno oko 15%. Udeo korisnika koji su napustili operatora u čitavom skupu podataka je 14.5%. Ovo ukazuje na to da *pozivni broj*, kao atribut, nema uticaja na vrednost ciljne varijable. Iz tog razloga je i ovaj atribut isključen, tj. obrisan iz okvira podataka.

Budući da mnogi algoritmi mašinskog učenja ne mogu da rade sa kategoričkim varijablama, već isključivo sa numeričkim, priprema skupa podataka često podrazumeva i kodiranje kategoričkih varijabli. Postoje dve vrste kodiranja kategoričkih varijabli: *Label Encoding* i *One-Hot Encoding*. *Label Encoding* podrazumeva zamenu kategoričkih vrednosti brojevima. *One-Hot Encoding* podrazumeva kreiranje nove binarne varijable za svaku moguću vrednost kategoričke varijable. Sada je jasno zašto upoznavanje sa skupom podataka obavezno treba da sadrži utvrđivanje broja prisutnih jedinstvenih vrednosti za svaku varijablu, a posebno za kategoričke. U izvornom skupu podataka

korišćenom u ovom istraživanju, nakon što su izbačene promenljive *broj telefona* i *drzava u SAD*, preostale su samo tri kategoričke varijable: *plan medjunarodnih poziva*, *mogucnost glasovne pošte* i ciljna varijabla *odliv korisnika*. Prve dve varijable uzimaju vrednosti *da* ili *ne* dok ciljna varijabla uzima vrednosti *TRUE* ili *FALSE*. Vrednosti *da* i *TRUE* kodirane su brojem 1, dok su vrednosti *ne* i *FALSE* kodirane brojem 0.



Slika 2. Procentualna raspodela odliva korisnika prema pozivnim brojevima

Zatim su od okvira podataka *df_telekomunikacije_odliv_korisnika_transformisan* kreirane *Python* promenljive `X` i `y`. Promenljiva `X` predstavlja dvodimenzionalnu kolekciju podataka koja sadrži vrednosti svih nezavisnih varijabli, dok promenljiva `y` predstavlja jednodimenzionalnu kolekciju podataka koja sadrži vrednosti zavisne varijable.

Skaliranje podataka je uobičajena praksa kod primene tehnike mašinskog učenja, a sastoji se od transformacije vrednosti iz numeričkih kolona u vrednosti na zajedničkoj skali. U mašinskom učenju, vrednosti nekih numeričkih atributa mogu biti više puta veće od vrednosti drugih atributa. Atributi koji imaju veće vrednosti će dominirati procesom učenja, iako to ne mora da znači da su te promenljive važnije u predviđanju vrednosti ciljne varijable. Normalizacija podataka je jedna od metoda koja transformiše podatke iz različitih skala u podatke na jednoj zajedničkoj skali. Nakon normalizacije, sve promenljive imaju sličan uticaj na model, što doprinosi poboljšanju stabilnosti i performansi modela mašinskog učenja. Postoji više tehnika normalizacije u statistici. U ovom istraživanju korišćena je min-max metoda normalizacije, koja sve numeričke vrednosti transformiše u vrednosti u fiksnom intervalu od 0 do 1. Ova transformacija se vrši tako što se od posmatrane vrednosti atributa oduzima minimalna vrednost tog atributa, a onda se ta razlika deli rasponom vrednosti tog atributa (maksimalna vrednost atributa – minimalna vrednost atributa). Veliki broj istraživača, koji koriste tehniku nadgledanog mašinskog učenja, preporučuju da se skaliranje vrednosti numeričkih varijabli radi nakon podele skupa podataka na skupove za trening i testiranje, odnosno na svakom od ovih skupova podataka odvojeno. Takođe, preporučuje se da se parametri normalizacije (minimalna i maksimalna vrednost svake numeričke varijable) odrede na skupu podataka za trening, a da se koriste u normalizaciji promenljivih oba skupa podataka (i za trening i za testiranje). Ovo istraživanje urađeno je u skladu sa pomenutim preporukama.

Pre nego što je izvršena normalizacija numeričkih varijabli, poslednja verzija skupa podataka, smeštena u promenljivim `X` i `y`, podeljena je na skup podataka za trening (75% instanci) i skup podataka za verifikaciju (testiranje) modela (25% instanci). Nakon toga, odvojeno je urađena normalizacija skupa podataka za trening i skupa podataka za testiranje.

2.3. Izgradnja modela mašinskog učenja

Nakon odgovarajuće pripreme skupova podataka za treniranje i testiranje modela mašinskog učenja, sproveden je proces mašinskog učenja, po sledeći fazama:

1. obučavanje različitih modela mašinskog učenja, na skupu podataka za trening,
2. validacija, tj. ocena obučenih modela,
3. izbor najboljeg modela za predikciju,
4. podešavanje vrednosti hiperparametara najboljeg modela,
5. verifikacija izabranog modela na skupu podataka za testiranje.

Na skupu podataka za trening obučeni su modeli bazirani na sledećim *Scikit-Learn* (*Sklearn*) klasifikatorima: *DummyClassifier*, *LogisticRegression*, *DecisionTreeClassifier*, *KNeighborsClassifier*, *LinearDiscriminantAnalysis*, *GaussianNB*, *SVC* i *RandomForestClassifier*. *Scikit-Learn* je najznačajnija *Python* biblioteka za mašinsko učenje. Prilikom obučavanja modela korišćene su podrazumevane vrednosti njihovih hiperparametara. *DummyClassifier* kreira predikcije ne uzimajući u obzir vrednosti nezavisnih atributa, već samo frekvencije pojavljivanja različitih vrednosti ciljne varijable u skupu podataka koji se koristi za obučavanje. Ovaj klasifikator je korišćen za obučavanje tzv. nultog modela, sa kojim će se porediti ostali obučeni modeli.

Za ocenu i poređenje obučenih modela izabrana je mera performansi koja se naziva tačnost modela *(Accuracy*). S obzirom na to da je izvorni skup podataka relativno mali (3333 instance), za validaciju modela, tj. izračunavanje performansi modela, korišćena je tehnika *K-fold cross-validation*, i to za K=10. Kod ove tehnike, skup podataka za trening deli se K puta na K podskupova. U svakoj od K iteracija jedan od K podskupova koristi se za izračunavanje tačnosti modela, a ostalih K-1 za obučavanje modela. Konačne tačnosti modela izračunavaju se kao aritmetičke sredine K dobijenih tačnosti, u K iteracija. Budući da je za validaciju modela u ovom istraživanju korišćena tehnika *10-fold cross-validation*, tačnost modela izračunata je korišćenjem funkcije *cross_val_score()* iz *Python sklearn.model_selection* modula.

Od osam obučenih modela, kao najbolji, izabran je onaj model koji je imao najveću tačnost. Tačnost najboljeg modela bila je značajno veća od tačnosti nultog modela, što je predstavljalo potvrdu da se predikcija odliva korisnika može raditi korišćenjem tehnike nadgledanog mašinskog učenja. Nakon toga, u cilju eventualnog poboljšanja tačnosti izabranog modela, izvršeno je podešavanje hiperparametara izabranog modela, korišćenjem tehnike *RandomizedSearchCV*.

U poslednjoj fazi izvršena je verifikacija najboljeg modela na skupu podataka za testiranje. Ova faza podrazumeva generisanje matrice konfuzije (*confusion matrix*) i izračunavanje različitih mera performansi modela, kao što su: tačnost, osetljivost, specifičnost i preciznost. Matrica konfuzije generiše se uz pomoć funkcije *confusion_matrix()* implementirane u *Python sklearn.metrics* modulu. Za problem

binarne klasifikacije sa nejednako zastupljenim klasama, ređe zastupljena klasa često nosi naziv pozitivna klasa, a zastupljenija klasa nosi naziv negativna klasa. U skupu podataka korišćenom u ovom istraživanju zastupljenija je klasa koja predstavlja korisnike koji nisu napustili operatora. Dakle, pozitivna klasa je ona klasa kod koje ciljna varijabla ima vrednost 1, a negativna je ona klasa kod koje ciljna varijabla ima vrednost 0 (Slika 3).

Ako matricu konfuzije za posmatrani problem binarne klasifikacije označimo sa C_{2x2}, tada će njeni elementi $c_{i,j}$ predstavljati broj instanci koje pripadaju klasi i a predikcijom su raspoređene u klasu j. Na taj način dobijamo sledeća značenja elemenata matrice konfuzije: $c_{0,0}$ – broj stvarno negativnih instanci (*true negative*, *TN*), $c_{0,1}$ – broj lažno pozitivnih instanci (*false positive*, *FP*), $c_{1,0}$ – broj lažno negativnih instanci (*false negative*, *FN*) i $c_{1,1}$ – broj stvarno pozitivnih instanci (*true positive*, *TP*).

		predviđene vrednosti ciljne varijable	
		0	1
stvarne vrednosti ciljne varijable	0	*TN*	*FP*
	1	*FN*	*TP*

Slika 3. Matrica konfuzije za rešavani problem binarne klasifikacije

Mere performansi klasifikacionih modela u ovom istraživanju izračunate su korišćenjem elemenata matrice konfuzije kao operanada u odgovarajućim izrazima. Tačnost (*Accuracy*) i osetljivost (*Sensitivity*) su izračunate prema (1), specifičnost (*Specificity*) i preciznost (*Precision*) prema (2), *F1 Score* prema (3):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad Sensitivity = \frac{TP}{TP+FN} \tag{1}$$

$$Specificity = \frac{TN}{TN+FP} \qquad Precision = \frac{TP}{TP+FP} \tag{2}$$

$$F1\ Score = \frac{2*Precision*Sensitivity}{Precision+Sensitivity} \tag{3}$$

3. Rezultati i analiza rezultata

Na skupu podataka za trening obučeno je osam modela, baziranih na različitim algoritmima mašinskog učenja. U Tabeli 5 prikazane su tačnosti svih obučenih modela. Lako se uočava da model *RandomForestClassifier* ima značajno veću tačnost od ostalih modela. Budući da nulti model (*DummyClassifier*) ima tačnost 0.850341, a najbolji model 0.950381, može se konstatovati da najbolji model ima značajno veću tačnost od nultog modela, i da ima smisla nastaviti proces mašinskog učenja korišćenjem ovog modela.

Tabela 5. Rezultati unakrsne validacije modela na skupu podataka za trening

Model mašinskog učenja	**Tačnost modela (*Accuracy*)**
DummyClassifier	0.850341
LogisticRegression	0.856347
LinearDiscriminantAnalysis	0.849949
KNeighborsClassifier	0.897952
DecisionTreeClassifier	0.908369
GaussianNB	0.861944
SVC	0.850341
RandomForestClassifier	0.950381

Nakon podešavanja hiperparametara najboljeg modela, izvršena je verifikacija ovog modela, tj. izračunavanje različitih mera njegovih performansi, na skupu podataka za testiranje. Matrica konfuzije najboljeg modela prikazana je na Slici 4, dok su u Tabeli 6 prikazane izračunate mere performansi ovog modela.

$$\begin{bmatrix} 722 & 3 \\ 30 & 79 \end{bmatrix}$$

Slika 4. Matrica konfuzije najboljeg modela na skupu podataka za testiranje

Iz matrice konfuzije može se lako izračunati, da je od ukupno 834 instance skupa podataka za testiranje, njih 33 pogrešno klasifikovano. Za 3 korisnika koji nisu napustili svog operatora, model je predvideo da će to učiniti, a za 30 korisnika koji jesu napustili operatora, model je predvideo da to neće učiniti. Model je tačno predvideo ponašanje za 722 korisnika koji nisu napustili operatora i za 79 korisnika koji su napustili operatora.

Tabela 6. Mere performansi najboljeg modela na skupu podataka za testiranje

Mera performansi	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	*F1 Score*
Vrednost mere	0.96	0.72	0.99	0.96	0.83

Tačnost modela na skupu podataka za testiranje se malo povećala u odnosu na tačnost izmerenu na skupu podataka za trening (sa 0.95 na 0.96), što znači da nije došlo do problema prevelikog podudaranja (engl. *Overfitting*). Tačnost modela je zadovoljavajuća, kao i vrednosti ostalih mera performansi, pa se može zaključiti da se model mašinskog učenja baziran na algoritmu *RandomForest*, kreiran u programskom jeziku *Python*, može uspešno koristiti u predikciji odliva korisnika telekomunikacionih operatora.

U literaturi su opisana brojna istraživanja koja su rezultovala modelima za predikciju odliva korisnika telekomunikacionih operatora. Kreirani modeli se međusobno razlikuju po tačnosti i korišćenom skupu podataka. Npr. u radu [4] korišćen je skup podataka koji sadrži zapise o 7032 korisnika, pri čemu je broj atributa u skupu podataka 21. Pored podataka o servisima koje su korisnici koristili, ovaj skup podataka sadrži i demografske podatke o korisnicima, poput pola i godišta. Tačnost najboljeg modela u ovom slučaju iznosi oko 82%. Prema [5] isti skup podataka korišćen je u velikom broju drugih radova poput [6, 7, 8], pri čemu se tačnost modela kreće u opsegu od 68% do

85%. Dostupnost podataka, koji su neophodni za kreiranje modela za predikciju odliva korisnika, je ograničena zbog poverljivosti i privatnosti podataka između telekomunikacionih operatora i njihovih korisnika [4]. Ovo predstavlja dodatnu prepreku prilikom kreiranja modela. Upravo iz tog razloga, u ovom istraživanju korišćen je otvoreni skup podataka. Tačnost najboljeg modela razvijenog u ovom istraživanju iznosila je preko 95%, što je za oko 10% više od tačnosti najboljih modela opisanih u pomenutim radovima.

4. Zaključak

Odliv korisnika direktno se odražava na uspešnost poslovanja telekomunikacionih operatora. Smanjenje odliva korisnika, pored povećanja profita i smanjenja troškova, može imati i pozitivan uticaj na privlačenje novih korisnika. Telekomunikacioni operatori sa manjom stopom odliva korisnika imaju zadovoljnije i lojalnije korisnike, što predstavlja veliku prednost na konkurentnom tržištu.

Prediktivni model kreiran u ovom istraživanju ima za oko 10% veću tačnost od sličnih modela koje su razvijali drugi istraživači, sa istim ciljem. Zbog visoke tačnosti, model razvijen u ovom istraživanju pogodan je za praktičnu implementaciju od strane telekomunikacionih operatora u cilju smanjenja odliva korisnika i povećanja profita. Osim toga, prediktivni model je obučen i testiran na skupu podataka koji ne sadrži osetljive kategorije podataka o korisnicima, kao što su ime i prezime, broj telefona, adresa i sl. Time je ispunjen još jedan važan preduslov za primenu ovog modela u praksi.

Zahvalnica

Ovaj rad delimično je podržan od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije.

Literatura

[1] N. N. Nguyen, and A. T. Duong, “Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem”, *Journal of advances in information technology*, vol. 12, no.1, pp. 29-35, February 2021. DOI: 10.12720/jait.12.1.29-35

[2] Orange. Available at: https://orangedatamining.com/

[3] Telecom_churn skup podataka. Available at: https://www.kaggle.com/datasets/keyush06/telecom-churncsv

[4] B. Yogesh, and R. T. Fokone, “Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry”, *Concurrency and Computation: Practice and Experience*, vol. 34, no. 3, September 2021. DOI: 10.1002/cpe.6627

[5] F. S. Wael, S. Subramanian, and M. A. Khder, “Customer Churn Prediction in Telecommunication Industry Using Deep Learning”, *Information Sciences Letters*, vol. 11, no. 1, pp. 185 – 198, January 2022. DOI: 10.18576/isl/110120

[6] S. Momin, T. Bohra, and P. Raut, “Prediction of Customer Churn Using Machine Learning”, *Proceedings of the EAI International Conference on Big Data Innovation*

for Sustainable Cognitive Computing, pp. 203–212, 2020. DOI: 10.1007/978-3-030-19562-5_20
[7] N. I. Mohammad, S. A. Ismail, M. N. Kama, O. M. Yusop, and A. Azmi, "Customer Churn Prediction in Telecommunication Industry Using Machine Learning Classifiers", *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing (ICVISP 2019)*, pp. 1–7, 2019. DOI: https://doi.org/10.1145/3387168.3387219
[8] J. Pamina, J. Beschi Raja, S. Sathya Bama, S. Soundarya, M. S. Sruthi, S. Kiruthika, V. J. Aiswaryadevi, and G. Priyanka, "An effective classifier for predicting churn in telecommunication", *Journal of Advanced Research in Dynamical and Control Systems*, vol. 11, no. 1, pp. 221-229, June 2019.

Abstract: *In order to prevent customer churn, it would be useful for the telecommunications operator to find out which parameters have the greatest influence on the users' churn. The paper deals with the problem of predicting future user churn, based on historical data, in the Python programming language. In order to solve this problem, a suitable, open data set was found and an exploratory data analysis was performed, in order to determine the degree of dependence between each independent and dependent variable. The independent variables describe the user and the services he/she used, while the dependent variable answers the question: has the user left the operator by that time? After that, different machine-learning classification models were created using some of the algorithms implemented in the Scikit-Learn library available in the Python programming language. The accuracy of the best models was over 95%, which is 10% more than the accuracy of the null model, so it can be concluded that the prediction of user churn can be successfully performed using machine learning, in the Python programming language.*

Keywords: *machine learning, churn, prediction, classification, Python*

PREDICTION OF CUSTOMER CHURN IN TELECOMMUNICATIONS USING MACHINE LEARNING

Slađana Janković, Snežana Mladenović, Ivana Stefanović, Ana Uzelac