https://doi.org/10.37528/FTTE/9788673955056/POSTEL.2025.20

UNAPREĐENJE PREDIKCIJE ODLIVA KORISNIKA TELEKOMUNIKACIONIH OPERATORA POMOĆU TEHNIKA BALANSIRANJA KLASA

Katarina Kukić, Ana Uzelac, Slađana Janković, Snežana Mladenović
Univerzitet u Beogradu - Saobraćajni fakultet,
k.kukic@sf.bg.ac.rs, ana.uzelac@sf.bg.ac.rs,
s.jankovic@sf.bg.ac.rs, snezanam@sf.bg.ac.rs

Rezime: *U radu je prikazano unapređenje predikcije odliva korisnika telekomunikacionih operatora primenom tehnike sintetičkog naduzorkovanja SMOTE. Korišćen je javno dostupan Orange telecom churn skup podataka sa izraženom neuravnoteženošću klasa (oko 14,5% odlazećih korisnika). Na neizbalansiranom i na SMOTE-om balansiranom skupu poređeni su klasični modeli mašinskog učenja (Decision Tree, k-Nearest Neighbors, Random Forest) i savremeni boosting algoritmi (XGBoost, LightGBM), a performanse su ocenjene metrikama accuracy, precision, recall i F1. Rezultati pokazuju da SMOTE značajno povećava recall za klasu odlazećih korisnika kod svih modela, dok boosting modeli i bez balansiranja postižu visoke performanse, a uz SMOTE dodatno unapređuju recall uz zadržavanje visoke tačnosti. SHAP analiza XGBoost modela otkriva da najveći uticaj na odlazak korisnika imaju broj poziva korisničkom servisu, ukupan broj potrošenih dnevnih i večernjih minuta, dostupnost međunarodnih razgovora i odsustvo glasovne pošte. Ovi nalazi potvrđuju da kombinacija SMOTE tehnike i naprednih modela omogućava pouzdaniju detekciju rizika odliva i identifikaciju ključnih faktora nezadovoljstva korisnika.*

Ključne reči: *SMOTE, odliv korisnika, mašinsko učenje, boosting algoritmi, SHAP analiza*

1. Uvod

U našem ranijem istraživanju [1] ispitivali smo predikciju odliva korisnika koristeći javno dostupan *dataset Orange telecom churn* [2]. Ovaj *dataset* sadrži 3333 zapisa o korisnicima i 21 atribut, od kojih približno 14,5 % predstavljaju zapise o korisnicima koji su prestali da koriste usluge operatora, što dovodi do izražene neuravnoteženosti klasa. U ranijem radu primenili smo nekoliko standardnih algoritama mašinskog učenja kao što su logistička regresija, *Decision Tree*, *k-Nearest Neighbors* i *Random Forest*. U fazi pripreme podataka sprovedeni su sledeći koraci: provera nedostajućih vrednosti (nije bilo potrebe za imputacijom), uklanjanje irelevantnih atributa, kodiranje kategoričkih promenljivih i podela skupa na trening (75 %) i test (25

%). Među primenjenim modelima *Random Forest* je postigao najbolje rezultate, sa tačnošću većom od 95 % i *F1*-merom od 0,83 na test skupu.

Iako su ovi rezultati ohrabrujući, u prethodnom istraživanju nije bila primenjena eksplicitna strategija za rešavanje neuravnoteženosti klasa, što može pristrasno da utiče na modele u korist većinske klase i da umanji njihovu sposobnost da pravilno identifikuju korisnike sklone odlasku.

Cilj ovog rada je prevashodno da prevaziđe ovo ograničenje primenom *Synthetic Minority Over-sampling Technique* (*SMOTE*). *SMOTE* je široko prihvaćen pristup na nivou podataka za balansiranje skupova kod binarne klasifikacije. Ova metoda generiše sintetičke primere manjinske klase interpolacijom između postojećih instanci manjinske klase u prostoru atributa, čime se stvara uravnoteženiji skup za obuku bez prostog dupliranja postojećih uzoraka. Brojne studije o predikciji odliva korisnika pokazuju efikasnost ove i sličnih tehnika [3]–[8]; dosledno izveštavaju o poboljšanom otkrivanju pripadnika manjinske klase i o višim vrednostima mera *recall* i *F1* kada se neuravnoteženost klasa tretira sintetičkim naduzorkovanjem. Napomenimo da ćemo, kao posebno značajnu metriku posmatrati vrednost metrike *recall* za manjinsku klasu odlazećih korisnika koja pokazuje koliko ih dobro model prepoznaje, jer ako je *recall* nizak, znači da model ne detektuje veliki broj korisnika koji odlaze – što je za kompaniju "najskuplji" propust, jer kompanija neće reagovati i ti korisnici će zaista otići. Ukoliko model pogrešno označi neke korisnike kao odlazeće (*false positive*), to će za kompaniju predstavljati manji gubitak nego propust da model dobro prepozna one koji zaista odlaze (*false negative*).

Na osnovu ovih nalaza, ponovo analiziramo isti *Orange* skup podataka kako bismo procenili da li balansiranje podataka pomoću *SMOTE*-a može dodatno da poboljša prediktivne performanse naših modela odliva, posebno u pravilnom identifikovanju potencijalnih korisnika koji bi mogli da napuste operatora. Osim toga, u ovom radu će, pored klasičnih modela koji su se u prethodnom istraživanju pokazali kao najtačniji (*Decision Tree*, *k-Nearest Neighbors* i *Random Forest*), biti primenjeni i savremeni *boosting* algoritmi *XGBoost* i *LightGBM*. *Boosting* metode su izabrane jer omogućavaju postizanje veće tačnosti i bolje generalizacije, posebno na kompleksnim i potencijalno neuravnoteženim skupovima podataka. Daćemo i interpretaciju dobijenih rezultata za model sa najboljim performansama, kao i SHAP analizu iz koje ćemo zaključiti o ukupnom uticaju svakog atributa na predikcije modela.

Ovaj rad je organizovan na sledeći način: u drugom poglavlju dat je pregled relevantne literature o problemu predikcije odliva korisnika i tehnikama za balansiranje klasa. Treće poglavlje opisuje korišćeni skup podataka i postupak preprocesiranja. U četvrtom poglavlju predstavljena je metodologija, uključujući primenjene modele mašinskog učenja i metrike evaluacije. Peto poglavlje donosi eksperimentalne rezultate i diskusiju, dok su u šestom poglavlju izneti zaključci i predlozi za buduća istraživanja.

2. Pregled literature

Odliv korisnika predstavlja visoko neuravnotežen binarni klasifikacioni zadatak, u kojem je manjinska klasa - korisnici koji odlaze – uobičajeno relativno mala. Jednostavni modeli mogu postići visoku ukupnu tačnost, ali i dalje mogu propustiti da prepoznaju korisnike sklone odlasku. Zbog toga mnoge studije eksplicitno razmatraju

problem neuravnoteženosti klasa kako bi poboljšale metrike manjinske klase, posebno *recall* i *F1*.

Najčešći pristupi su ponovo uzorkovanje na nivou podataka – na primer *SMOTE* i njegove varijante – dok se ređe primenjuju pažljivo dizajnirane strategije poduzorkovanja. U radu [3] upoređeno je šest metoda naduzorkovanja, uključujući *SMOTE* i *ADASYN*, i pokazano je da ponovno uzorkovanje može značajno da poboljša detekciju manjinske klase, iako veličina dobitka zavisi od konkretne metode. Korak dalje napravljen je u radu [4], gde se, pored upoređivanja metoda naduzorkovanja (*SMOTE* i *ADASYN*), uvodi novi dvostepeni pristup: najpre se primenjuje klasterovanje radi uklanjanja šuma, a zatim se koristi ansambl metoda za generisanje uzoraka manjinske klase.

U [9] je predložen integrisani okvir koji povezuje predikciju odliva sa segmentacijom korisnika. Nakon što *SMOTE* poboljša ravnotežu klasa u fazi predikcije, primenjuje se Bajesova logistička regresija s ciljem da se identifikuju glavni faktori odliva, a potom se ti faktori koriste za *K-means* klasterovanje korisnika. Na ovaj način pokazano je da kombinovanje predikcije i segmentacije vodi ka konkretnijim i primenljivijim strategijama zadržavanja korisnika. Prednosti primene SMOTE tehnike potvrđene su i u drugim radovima [6, 7, 10]. U [11] je analizirano jedanaest klasifikatora i sedam pristupa uzorkovanju na šesnaest skupova podataka sličnih onima koji se koriste za predikciju odliva. Rezultati pokazuju da uzorkovanje može da poveća vrednosti metrika *recall* i *F1*, ali i da smanji performanse jakih klasifikatora kao što je *Random Forest* kada karakteristike skupa već idu u korist ovih modela, što ukazuje na potrebu pažljive evaluacije.

Pored naduzorkovanja, neka istraživanja beleže i različite pristupe poduzorkovanju [12, 13, 14], ali oni nisu predmet ovog rada. Takođe, neka istraživanja kombinuju naduzorkovanje i poduzorkovanje: na primer, u [8] primenjen je *ADASYN*, a zatim *NEASMISS* poduzorkovanje, čime je značajno poboljšan *recall* na višemilionskom *dataset*-u, dok je u [15] pokazano da *SMOTE* daje bolje rezultate od čistog poduzorkovanja.

Sve navedene studije ponovno uzorkovanje primenjuju isključivo na trening skupu, čime se izbegava curenje podataka (engl. *data leakage*) – isti protokol sledi i naše istraživanje. Velike komparativne studije potvrđuju da se efikasnost uzorkovanja razlikuje u zavisnosti od modela i samog skupa podataka. Ovo naglašava potrebu za kontrolisanim empirijskim poređenjima kakvo i mi predstavljamo. Na osnovu ovih nalaza ponovo analiziramo skup podataka *Orange telecom churn* iz našeg ranijeg rada i sistematski ispitujemo primenu *SMOTE* metode u istom okruženju mašinskog učenja. Cilj je da kvantifikujemo uticaj *SMOTE*-a na metrike detekcije manjinske klase, pri čemu se ponovno uzorkovanje, u skladu s dobrom praksom, primenjuje samo na trening skup radi fer poređenja.

3. Dataset i preprocesiranje

Za potrebe analize korišćen je javno dostupan *Orange telecom churn* skup podataka koji obuhvata 3333 zapisa i ukupno 21 atribut. Ciljna promenljiva „odliv korisnika“ (*churn*) označava da li je korisnik napustio operatora (1 – „Otišao“, 0 – „Ostao“). Da bi skup podataka bio pogodan za primenu algoritama mašinskog učenja, izvršena je transformacija u potpuno numerički *dataset* kroz sledeće korake:

1) Uklonjeni su atributi država u SAD, pozivni broj i broj telefona, jer nemaju uticaj na predikciju odliva.
2) Kategoričke promenljive plan međunarodnih poziva i mogućnost glasovne pošte kodirane su binarno (da = 1, ne = 0), a ciljna varijabla je takođe kodirana u formatu 1/0.

Nakon uklanjanja irelevantnih kolona i kodiranja kategoričkih promenljivih, kreirane su matrica atributa (**X**) i ciljna promenljiva (**y**). U numeričkom datasetu su zadržani sledeći atributi: *trajanje računa u danima*, *plan međunarodnih poziva (0/1)*, *mogućnost glasovne pošte (0/1)*, *broj glasovnih poruka*, *ukupan broj minuta dnevno*, *ukupan broj poziva dnevno*, *ukupna cena dnevnih poziva*, *ukupan broj večernjih minuta*, *ukupan broj večernjih poziva*, *ukupna cena večernjih poziva*, *ukupan broj noćnih minuta*, *ukupan broj noćnih poziva*, *ukupna cena noćnih poziva*, *ukupan broj međunarodnih minuta*, *ukupan broj međunarodnih poziva*, *ukupna cena međunarodnih poziva*, *broj poziva korisničkom servisu***.**

Korišćeni *dataset* potiče iz ranijeg perioda (oko 1998–1999), pa pojedini atributi – poput broja glasovnih poruka ili međunarodnog plana – danas mogu biti prevaziđeni u savremenim telekom mrežama. Iako je skup i dalje koristan za metodološka poređenja i evaluaciju algoritama, rezultati bi bili zanimljiviji i praktično relevantniji kada bi se primenio noviji, lokalno prikupljen skup podataka.

Skup je podeljen na trening (75%) i test (25%) deo, a sve numeričke promenljive normalizovane su *min–max* metodom na opseg [0, 1] radi ujednačavanja skala i sprečavanja dominacije atributa sa većim apsolutnim vrednostima.

SMOTE je tehnika balansiranja podataka koja se koristi kada imamo neravnotežu klasa (kao što je slučaj u razmatranom skupu podataka gde klasa odlazećih korisnika ima udeo od oko 14.5% u celom skupu podataka). Ova tehnika ne duplira postojeće primere, već stvara nove, sintetičke uzorke na osnovu postojećih, što vodi do realističnijeg proširenja manjinske klase. *SMOTE* se sprovodi tako što se za svaki uzorak iz manjinske klase, traže njegovi najbliže susedi u istoj klasi (obično k=5), a zatim se vrši interpolacija – nasumično bira jednog od tih suseda i pravi novi sintetički uzorak tako što ga postavlja negde između postojećeg primera i izabranog suseda. Proces se ponavlja dok manjinska klasa ne dostigne željenu veličinu (najčešće jednaku većinskoj). Na taj način dobija se balansiran *dataset* u kome su klase približno podjednako zastupljene, što omogućava da modeli budu obučeni na ravnotežnim podacima, bolje prepoznaju manjinsku klasu i obezbede pouzdaniju interpretaciju rezultata.

Dodatno, za interpretaciju modela primenjena je i *SHAP* (engl. *SHapley Additive Explanations*) analiza. *SHAP* metoda se zasniva na Šeplijevim vrednostima [16] iz kooperativne teorije igara i omogućava fer i konzistentno kvantifikovanje doprinosa svakog obeležja predikciji modela [17].

4. Metodologija

Polazeći od rezultata dobijenih u prethodnom radu [1], odlučili smo se da odaberemo tri modela sa najboljim performansama među osam testiranih osnovnih modela mašinskog učenja. Pored njih najpre ćemo na neizbalansiranim, a zatim i na podacima balansiranim pomoću *SMOTE* tehnike primeniti još dva *boosting* modela:

XGBoost poznat po svojoj robusnosti i visokim performansama, kao i *LightGBM* koji obezbeđuje bržu i efikasniju obuku modela.

Pošto je *Orange telecom churn* skup podataka izrazito neuravnotežen, oslanjanje samo na ukupnu tačnost može biti varljivo: klasifikator koji bi sve korisnike predvideo kao one koji ne odlaze mogao bi da postigne visoku tačnost, a da pritom uopšte ne detektuje one koji odlaze. Zato, da bismo dobili uravnoteženiju sliku performansi modela, koristimo više komplementarnih metrika koje se uobičajeno preporučuju za neuravnotežene binarne klasifikacije:

Tačnost (engl. *accuracy*) – udeo ispravno klasifikovanih instanci u odnosu na ukupan broj uzoraka (1). Iako daje osnovnu informaciju o ukupnoj uspešnosti modela, tačnost treba tumačiti oprezno jer je snažno pod uticajem većinske klase i sama po sebi ne pokazuje koliko dobro model otkriva korisnike sklone odlasku.

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

Preciznost (engl. *precision*) – udeo korisnika koje je model predvideo kao korisnike koji žele da odu, a koji su zaista otišli (2). Visoka preciznost znači da, kada model predvidi odlazak korisnika, ta prognoza je uglavnom tačna.

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

Osetljivost (engl. *recall*) – udeo stvarnih korisnika koji napuštaju operatora koje je model ispravno identifikovao (3). Za upravljanje odlivom *recall* je presudan, jer njegovo nedetektovanje znači gubitak korisnika koji je mogao da bude zadržan. Zato se posebno fokusiramo na vrednost *recall* metrike za klasu korisnika koji odlaze.

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \tag{3}$$

F1-mera (engl. *F1-score*) – harmonijska sredina preciznosti i osetljivosti (4). Ova jedinstvena metrika sažima odnos između otkrivanja što više stvarnih korisnika koji napuštaju operatora (visok *recall*) i izbegavanja lažnih alarma (visoka preciznost), pa daje uravnotežen prikaz performansi modela.

$$F1\ Score = \frac{2*Precision*Sensitivity}{Precision+Recall} \tag{4}$$

U prethodnom radu koristili smo i meru specifičnost (engl. *specificity*), koja pokazuje koliko dobro model prepoznaje korisnike koji ostaju. Međutim, u ovom istraživanju fokus je na pravovremenom otkrivanju onih koji odlaze i poboljšanju detekcije manjinske klase. Zbog toga se sada naglasak stavlja na metrike osetljive na neuravnoteženost – pre svega *recall* i *F1*-meru – dok je specifičnost, iako i dalje informativna, za ovu analizu sekundarnog značaja i nije uključena u glavni skup metrika.

Napomenimo još da se performanse modela treniranih uz *SMOTE* tehniku uvek procenjuju na neizmenjenom test skupu, jer je cilj da se proveri kako se model generalizuje na realne, izvorne podatke. Ako bismo *SMOTE* primenili i na test skupu, došlo bi do veštačkog balansiranja i rezultati bi bili nerealno optimistični, što ne bi odgovaralo stvarnim uslovima. Na taj način obezbeđujemo da evaluacija modela verno odražava sposobnost predviđanja odliva korisnika u praksi.

5. Rezultati i diskusija

Napomenimo da prikazani rezultati verovatno nisu najbolji mogući jer u razmatranim modelima nije rađena optimizacija hiperparametara, iz razloga što je primarni cilj ovog rada bilo da se pokaže kako balansiranje klasa u podacima doprinosi poboljšanju performansi modela. U tom cilju hteli smo da modeli razmatrani sa i bez primene SMOTE tehnike budu uporedivi, dok bi optimizacija hiperparametara verovatno podrazumevala drugačije vrednosti hiperparametara u ova dva pristupa.

Rezultati svih primenjenih modela, sa i bez SMOTE tehnike balansiranja klasa, prikazani su u Tabeli 1:

Tabela 1: Performanse modela sa i bez primene SMOTE tehnike

Model	SMOTE	Accuracy (%)	Klase	Precision (%)	Recall (%)	F1-score (%)
KNN	Ne	88	Ostao	89	98	93
			Otišao	70	26	38
	Da	67	Ostao	90	69	78
			Otišao	23	56	33
DT	Ne	93	Ostao	94	98	96
			Otišao	85	66	74
	Da	92	Ostao	95	96	96
			Otišao	75	71	73
RF	Ne	91	Ostao	91	100	95
			Otišao	96	39	55
	Da	90	Ostao	91	98	94
			Otišao	77	46	58
XGBoost	Ne	94	Ostao	95	98	97
			Otišao	86	73	79
	Da	95	Ostao	96	98	97
			Otišao	85	78	81
LightGBM	Ne	95	Ostao	96	98	97
			Otišao	87	74	80
	Da	94	Ostao	96	97	97
			Otišao	81	78	79

Analiza prikazanih rezultata u Tabeli 1 jasno pokazuje značaj primene *SMOTE* tehnike za balansiranje podataka u zadatku predikcije odliva korisnika. Kod modela treniranih na originalnom (neizbalansiranom) skupu podataka, uočava se visoka tačnost i odlični rezultati za klasu „Ostao“, dok je za klasu „Otišao“ *recall* izrazito nizak (npr. 26% kod *KNN* i 39% kod *RF*). Ovo znači da modeli često ne prepoznaju korisnike koji napuštaju kompaniju, iako je upravo ova klasa od ključnog značaja za donošenje poslovnih odluka. Primetimo i da *boosting* modeli *XGBoost* i *LightGBM* koji nisu bili primenjivani u prethodnom istraživanju daju značajno bolje rezultate i na neizbalansiranom skupu podataka, posebno po pitanju *recall* metrike, što ih čini svakako prvim izborom za dalju primenu i interpretaciju dobijenih rezultata. Napomenimo da je od primenjenih standardnih modela mašinskog učenja, stablo odlučivanja (*DT*) pokazalo

najbolje performanse, ali treba imati na umu da ovaj model ima tendenciju preprilagođavanja podacima, pa bi zahtevalo dodatno ispitivanje i izvesan oprez pri interpretaciji rezultata.

Primena *SMOTE* tehnike dovodi do poboljšanja *recall* vrednosti za klasu „Otišao*"* kod svih modela. Posebno se ističe *KNN*, gde *recall* raste sa 26% na 56%, i *RF*, gde raste sa 39% na 46%. Time modeli uspešnije detektuju korisnike koji odlaze, što je u skladu sa ciljem istraživanja. Međutim, povećanje *recall*-a često dolazi po cenu smanjenja ukupne tačnosti (*accuracy*), jer se modeli suočavaju sa sintetički proširenim podacima i balansiranjem. Na primer, kod *KNN* tačnost opada sa 88% na 67%. Ipak, kod naprednijih modela kao što su *XGBoost* i *LightGBM*, primena *SMOTE* tehnike dovodi do balansiranijeg učinka, gde se *recall* za klasu „Otišao" dodatno povećava (*XGBoost* sa 73% na 78%, *LightGBM* sa 74% na 78%), dok tačnost ostaje visoka (94–95%). Ovi modeli uspešno integrišu informacije iz proširenog skupa i pokazuju najbolji kompromis između detekcije odliva i ukupne tačnosti.

Da sumiramo, i na neizbalansiranim i na balansiranim podacima najbolje performanse pokazali su *boosting* modeli, ali je primena *SMOTE* tehnike poboljšala i njihove performanse, posebno po pitanju vrednosti *recall* metrike. U nastavku ćemo se fokusirati na *XGBoost* model sa primenom *SMOTE* tehnike, i dati interpretaciju zaključaka koji se mogu izvući iz modela.

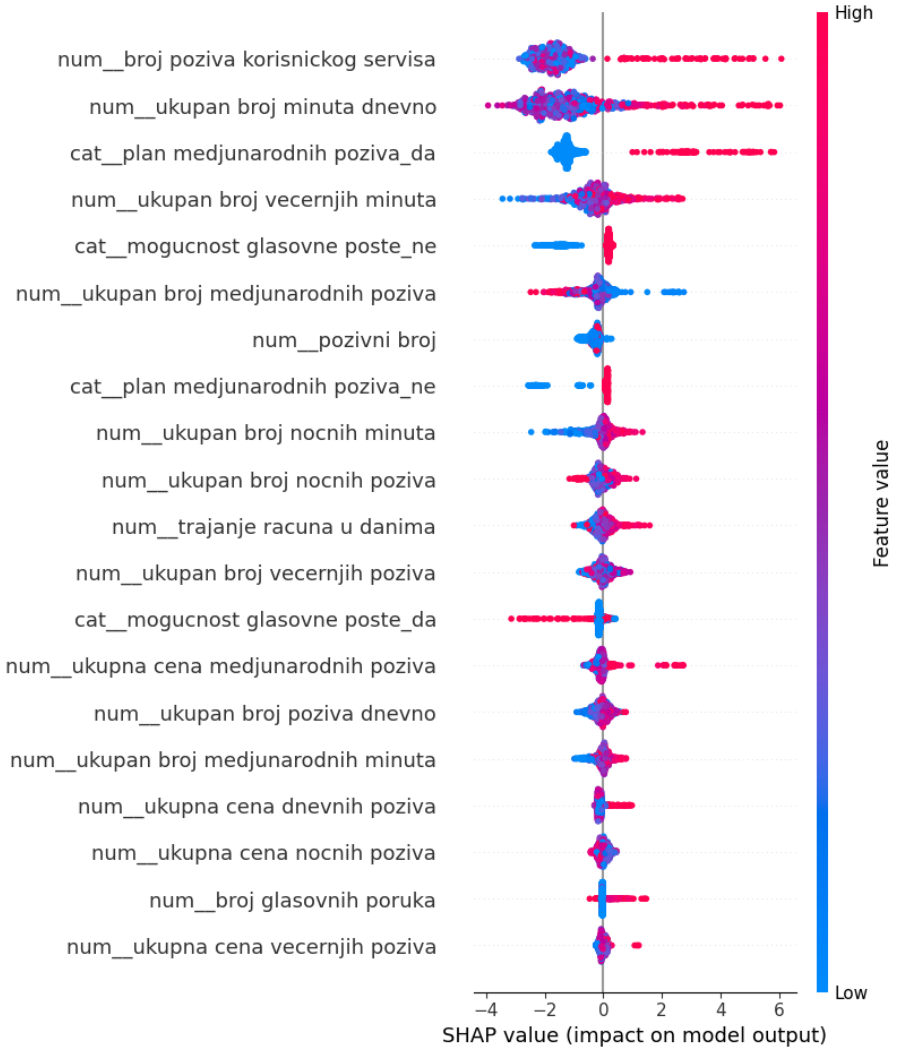
XGBoost model računa koliko puta i koliko efikasno je neki atribut korišćen za smanjenje greške i tako određuje važnost pojedinih atributa (engl. *feature importance*) – ako je atribut često korišćen za razdvajanje podataka i pritom značajno smanjuje grešku klasifikacije, njegova važnost biće visoka. Pored određivanja značaja atributa na ovaj način, moguće je izračunati i *SHAP* vrednosti. One ne samo da daju kvantitativnu meru doprinosa atributa predikciji, već kvalitativno pokazuju kako atribut utiče na izlaz modela: za svaki uzorak u skupu podataka određuje se da li atribut povećava ili smanjuje verovatnoću pripadnosti posmatranoj klasi i koliki je intenzitet tog uticaja. Ova tehnika obezbeđuje lokalnu tačnost, nulti doprinos obeležja bez uticaja i konzistentnost pri promeni značaja, pa se smatra jednom od najpouzdanijih metoda za interpretaciju kompleksnih modela. Njena primena omogućava i globalno sagledavanje značaja atributa i objašnjenje pojedinačnih predikcija.

Na Slici 1 prikazane su *SHAP* vrednosti za sve atribute, gde svaka tačka predstavlja jedan uzorak, intenzitet boje prikazuje vrednost *SHAP* vrednosti za posmatrani atribut, a položaj po horizontalnoj osi pomeren udesno ukazuje da atribut povećava verovatnoću da će uzorak pripadati klasi „Otišao". Pet atributa koji najviše doprinose odluci redom su: *broj poziva korisničkom servisu*, *ukupan broj minuta dnevno, plan međunarodnih poziva (da), ukupan broj večernjih minuta* i *mogućnost glasovne pošte (ne)*.

Detaljnija interpretacija rezultata prikazanih na Slici 1 ukazuje da atribut *Broj poziva korisničkom servisu* dominantno prikazan crvenim tačkama (mnogo poziva) koje se nalaze desno ukazuju da veliki broj poziva korisničkom servisu povećava šansu da korisnik ode. Slično, visoke vrednosti atributa *Ukupan broj minuta dnevno* (crvene tačke) takođe podstiču korisnika na odluku ka odlasku od operatora, dok niže vrednosti (plavo) podstiču odluku korisnika ka ostanku kod operatora.

Dalje, prisustvo plana međunarodnih poziva povećava šansu odlaska korisnika od operatora, dok njegovo odsustvo (plavo) znači veću verovatnoću ostanka. Po pitanju atributa *Večernji minuti*, vidi se da visoke vrednosti utiču da ishod predikcije bude

odlazak od operatora, što može ukazivati na specifičan tip potrošačkog ponašanja. Primećuje se i da su korisnici koji ne poseduju glasovnu poštu (crvene tačke) nešto češće vezani za odlazak od operatora.



Slika 1: SHAP dijagram značaja atributa u XGBoost modelu

6. Zaključak

Boosting algoritmi su se pokazali kao najefikasniji pristup za predikciju odliva korisnika, jer čak i bez balansiranja podataka postižu visoku tačnost i pouzdano identifikuju ključne atribute. Ipak, izrazita neuravnoteženost klasa – sa značajno manjim brojem korisnika koji napuštaju operatora – zahteva primenu tehnika kao što je *SMOTE*.

Naši rezultati pokazuju da *SMOTE* značajno povećava *recall* za manjinsku klasu, što je u praksi presudno jer smanjuje rizik da potencijalni odlazeći korisnici ostanu neprepoznati.

Dodatno, *SHAP* analiza pruža interpretabilnost modela i omogućava da se precizno identifikuju faktori rizika za odlazak korisnika. Veliki broj poziva korisničkom servisu, visok ukupan broj dnevnih i večernjih minuta, prisustvo međunarodnog plana i odsustvo glasovne pošte izdvajaju se kao najuticajniji prediktori odliva. Ovi nalazi mogu služiti kao osnova za proaktivne poslovne strategije – pravovremene mere za zadržavanje korisnika i unapređenje kvaliteta usluga.

Kombinovanjem *SMOTE* tehnike sa savremenim *boosting* algoritmima i interpretacijom rezultata putem *SHAP* analize, obezbeđuje se pouzdan i transparentan okvir za predikciju odliva. Buduća istraživanja trebalo bi da uključe novije i lokalno prikupljene skupove podataka kako bi se dodatno potvrdila generalizacija ovih nalaza i procenila primenljivost predloženog pristupa u savremenim telekomunikacionim okruženjima.

Literatura

[1] S. Janković, S. Mladenović, I. Stefanović, and A. Uzelac, "Predikcija odliva korisnika telekomunikacionih operatora primenom mašinskog učenja," in *Proc. PosTel*, Beograd, Srbija, Nov. 2022, pp. 249–258.

[2] Orange Telecom Churn Dataset. [Online]. Available: https://www.kaggle.com/datasets/mnassrib/telecom-churn-datasets

[3] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access*, vol. 4, pp. 7940–7957, Oct. 2016. DOI: 10.1109/ACCESS.2016.2619719

[4] S. J. Haddadi, A. Farshidvard, F. D. S. Silva, J. C. dos Reis, and M. da Silva Reis, "Customer churn prediction in imbalanced datasets with resampling methods: A comparative study," *Expert Syst. Appl.*, vol. 246, Art. no. 123086, Jul. 2024. DOI: 10.1016/j.eswa.2024.123086

[5] C. Rao, Y. Xu, X. Xiao, F. Hu, and M. Goh, "Imbalanced customer churn classification using a new multi-strategy collaborative processing method," *Expert Syst. Appl.*, vol. 247, Art. no. 123251, Aug. 2024. DOI: 10.1016/j.eswa.2024.123251

[6] S. Saha, C. Saha, M. M. Haque, M. G. R. Alam, and A. Talukder, "ChurnNet: Deep learning enhanced customer churn prediction in telecommunication industry," *IEEE Access*, vol. 12, pp. 4471–4484, 2024. DOI: 10.1109/ACCESS.2024.3352221

[7] O. Soleiman-garmabaki and M. H. Rezvani, "Ensemble classification using balanced data to predict customer churn: A case study on the telecom industry," *Multimedia Tools Appl.*, vol. 83, no. 15, pp. 44799–44831, Oct. 2023. DOI: 10.1007/s11042-023-15264-9

[8] J. B. G. Brito, G. B. Bucco, R. Heldt, J. L. Becker, C. S. Silveira, F. B. Luce, and M. J. Anzanello, "A framework to improve churn prediction performance in retail banking," *Financial Innovation*, vol. 10, no. 1, pp. 1–29, Jan. 2024. DOI: 10.1186/s40854-023-00558-3

[9] S. Wu, W.-C. Yau, T.-S. Ong, and S.-C. Chong, "Integrated churn prediction and customer segmentation framework for Telco business," *IEEE Access*, vol. 9, pp. 62118–62136, 2021. DOI: 10.1109/ACCESS.2021.3073776

[10] P. Jiang, Z. Liu, L. Zhang, and J. Wang, “Hybrid model for profit-driven churn prediction based on cost minimization and return maximization,” *Expert Syst. Appl.*, vol. 228, Art. no. 120354, Oct. 2023. DOI: 10.1016/j.eswa.2023.120354
[11] L. Geiler, S. Affeldt, and M. Nadif, “A survey on machine learning methods for churn prediction,” *Int. J. Data Sci. Analytics*, vol. 14, no. 3, pp. 217–242, Sep. 2022. DOI: 10.1007/s41060-021-00274-7
[12] K. Eria and B. P. Marikannan, “Systematic review of customer churn prediction in the telecom sector,” *J. Appl. Technol. Innov.*, vol. 2, no. 1, pp. 1–14, 2018.
[13] N. Edwine, W. Wang, W. Song, and D. Ssebuggwawo, “Detecting the risk of customer churn in telecom sector: A comparative study,” *Math. Probl. Eng.*, vol. 2022, pp. 1–16, Jul. 2022. DOI: 10.1155/2022/8115931
[14] A. Idris, A. Iftikhar, and Z. U. Rehman, “Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling,” *Cluster Comput.*, vol. 22, no. S3, pp. 7241–7255, May 2019. DOI: 10.1007/s10586-018-1849-7
[15] W. H. Khoh, Y. H. Pang, S. Y. Ooi, L.-Y.-K. Wang, and Q. W. Poh, “Predictive churn modeling for sustainable business in the telecommunication industry: Optimized weighted ensemble machine learning,” *Sustainability*, vol. 15, no. 11, p. 8631, May 2023. DOI: 10.3390/su15118631
[16] L. S. Shapley, “A value for n-person games,” in *Contributions to the Theory of Games*, vol. II, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307–317.
[17] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017. Available: https://arxiv.org/abs/1705.07874

Abstract: *This paper explores improving telecom customer churn prediction with the Synthetic Minority Oversampling Technique (SMOTE). The publicly available Orange telecom churn dataset, which has a strong class imbalance (about 14.5% churners), was used. Classical machine-learning models (Decision Tree, k-Nearest Neighbors, Random Forest) and modern boosting algorithms (XGBoost, LightGBM) were evaluated on both the original and SMOTE-balanced data using accuracy, precision, recall, and F1 metrics. SMOTE markedly increases recall for the churn class in all models, while boosting algorithms perform well even without balancing and, when combined with SMOTE, further improve recall while maintaining high accuracy. SHAP analysis of the XGBoost model highlights the strongest churn drivers: number of customer-service calls, total daytime and evening minutes used, presence of an international calling plan, and lack of voicemail. These results show that combining SMOTE with advanced models enables more reliable churn detection and reveals key factors of customer dissatisfaction.*

Keywords: *SMOTE, customer churn, machine learning, boosting algorithms, SHAP analysis*

IMPROVING TELECOM CUSTOMER CHURN PREDICTION USING CLASS BALANCING TECHNIQUES

Katarina Kukić, Ana Uzelac, Slađana Janković, Snežana Mladenović