Razvoj Data mining modela za predikciju prelaska Prepaid korisnika u telekomu

Marin Mandić, Goran Kraljević, Ivan Boban
Fakultet strojarstva i računarstva
Sveučilište u Mostaru
Mostar, Bosna i Hercegovina
marin.mandic@hteronet.ba, goran.kraljevic@hteronet.ba, ivan.boban1@hteronet.ba

Sažetak—**Zbog velike konkurencije na tržištu, telekom operateri su pogođeni prelaskom korisnika (engl. churn), stoga im je jako važno prepoznati koji korisnici će ih vjerojatno napustiti i preći u konkurentnog telekoma. Ovo istraživanje koristi podatke o ponašanju korisnika iz telekom sustava koji služe da bi se prepoznalo uzorke u ponašanjima i na takav način prepoznalo prelaske korisnika. Kod pripreme podataka napravljena je selekcija korisnih atributa korištenjem analize glavnih komponenti (eng. *Principal Component Analysis - PCA*). Također je napravljena i normalizacija vrijednosti atributa, da bi se postigla pravilna ravnoteža utjecaja svih atributa. U radu je kreirano više prediktivnih modela za detekciju churna Prepaid korisnika u telekomu i izvršena je analiza uspješnosti implementiranih Data mining modela.**

Ključne riječi-prediktivno modeliranje, Prepaid, Data mining, algoritmi strojnog učenja, churn

I. UVOD

Velika konkurencija na telekom tržištu prouzrokuje prelaske korisnika (eng. *churn*), gdje takvi korisnici prestaju koristiti usluge jednog telekoma i prelaze u konkurentni telekom. Prelazak korisnika za telekom uzrokuje gubitak prihoda, i stoga je telekomu jako važno utjecati na smanjenje takvih prelazaka. Da bi spriječili sam prelazak korisnika, potrebno je prepoznati koji korisnici će, sa nekom vjerojatnošću, preći u konkurentni telekom. Proces predikcije koristi prepoznavanje uzoraka u ponašanju korisnika da bi predvidio njegovo buduće ponašanje.

Telekomi posjeduju povijesne podatke o ponašanju svojih korisnika pa ih mogu upotrijebiti u procesu predikcije prelaska korisnika. Ako telekom uspije predvidjeti koji korisnik će ga napustiti onda ga može pokušati zadržati sa određenim marketinškim kampanjama. Marketinške kampanje su aktivnosti usmjerene prema korisniku u cilju povećavanja zadovoljstva korisnika. Na ovakav način je moguće zadržati postojeće korisnike što će prouzrokovati povećanje tj. očuvanje prihoda i dobiti za telekom.

V. Lazarov i M. Capota u svom radu [1] navode da je pet do šest puta skuplje dovesti novog korisnika nego zadržati postojećeg. To je još jedan razlog zašto je bitno zadržati korisnike i prepoznati one sa potencijalom prelaska u konkurenti telekom.

Ovaj rad je organiziran na način da su u poglavlju II opisana dosadašnja istraživanja u konkretnom području, u poglavlju III predstavljen je kreirani model za predikciju prelazaka korisnika, definicija churn-a, eksperiment i rezultati eksperimenta. Na kraju, u poglavlju IV prikazan je zaključak rada i u poglavlju V pravci daljnjeg istraživanja.

II. DOSADAŠNJA ISTRAŽIVANJA

Napravljeno je jako puno istraživanja na području predviđanja prelaska kupaca u telekom industriji. U radu je predstavljen kratki pregled dosadašnjih istraživanja na ovom području u posljednjih nekoliko godina.

Kiran Dahiya i Surbhi Bhatia u svom radu [2] su definirali pet koraka procesa predikcije prelazaka korisnika, a to su:

a) prikupljanje podataka,
b) priprema podataka,
c) predprocesiranje podataka,
d) ekstrakcija varijabli,
e) primjena algoritma strojnog učenja.

Točno definiran proces pomaže u rješavanju kompleksnog problema predikcije prelaska kupaca.

Grupa autora je radu [3] napravila usporedbu metoda naduzorkovanja sa primjenom na problem predikcije prelaska kupaca. Usporedili su metode MTDF, SMOTE, ADASYN, TRkNN, MWMOTE i COTE. Rezultati njihovog istraživanja su pokazali da je nabolje koristiti metodu MTDF. Ova metoda pomože u rješavanju problema neravnoteže u podacima, koja je uobičajena kod predikcije prelaska kupaca u telekomu. Problem neravnoteže je u tome što imamo značajno veći broj neprelaznika u odnosu na prelaznike.

Hui Li, Deliang i drugi autori su u svom radu [4] opisali važnost selekcije atributa za izgradnju modela predikcije. Selekcija atributa poboljšava model na tri načina: pojednostavljuje model pa ga je jednostavnije interpretirati, skraćuje vrijeme treniranja modela i poboljšava generalizaciju i na takav način smanjuje pretjeranu prilagođenost modela podacima za treniranje. Oni su u svom radu koristili *random forest* za mjerenje važnosti pojedine varijable.

Za izgradnju uspješnog modela za predikciju prelaska korisnika moguće je koristiti različite metode. Najčešće korištene metode su stabla odlučivanja [2], [4], logistička regresija [2], [5] i neuronska mreža [5], [6]. Pored gore navedenih metoda uspješno se koriste Bayesove mreže [7], *Support vector machines* (SVM) [8] i različiti hibridni modeli [9]. Uspješnost ovih modela jako ovisi o setu podataka nad kojima se model trenira.

Backiel.Y. Verbinnen, B. Baesens i G. Claeskens su u svom radu [10] primijenili socijalnu mrežu, kao dodatni podatak koji povećava točnost predikcije churn-a. Analiza socijalnih mreža se zasniva na socijalnom utjecaju jednih osoba na druge. Osoba koja je churn ima utjecaja na osobe iz svog socijalnog kruga. Dokazali su da podaci o socijalnim mrežama poboljšavaju točnost na način da su napravili usporedbu dva modela, jedan koji koristi podatke o socijalnim mrežama i model koji ne koristi podatke o socijalnim mrežama.

III. KREIRANJE MODELA ZA PREDIKCIJU PRELAZAKA KORISNIKA

A. Ulazni podaci

Telekomi posjeduju jako puno podataka o svojim korisnicima jer sakupljaju podatke o njihovom ponašanju. Velika količina tih podataka može biti redundantna i prouzrokovati manju točnost modela. Također, velika količina podataka uzrokuje i sporost izvršavanja modela.
Podaci koji se koriste u ovom radu mogu se podijeliti na:

1. Podaci o ponašanju korisnika koji sadrže atribute kao što su trajanje poziva unutar mreže, trajanje poziva prema drugim mobilnim mrežama, broj poslanih SMS-ova, količina prenesenih podataka i još 50-tak drugih, koji se tiču ponašanja korisnika.

2. Podaci o samim korisnicima, kojih u ovom radu nema puno, jer se radi o prepaid (anonimnim) korisnicima, pa nemamo podataka kao što su: spol, godine, bračno stanje itd. Za ovaj rad koriste se atribut vjernosti samog korisnika tj. vremena od kada korisnik koristi usluge telekoma i broj dana do isteka statusa „aktivnog korisnika“.

Broj podataka koji se koriste za eksperiment prikazani su u tabeli 1.

TABELA I. ULAZNI PODACI EKSPERIMENTA

ULAZNI PODACI	
Broj atributa	109
Broj zapisa/korisnika	49.868

Broj korisnika koji su označeni kao prelaznici je 6328, što znači da imamo 12,69% prelaznika.
Uobičajen problem u podacima za predikciju je disbalans, uvažavajući ciljanu varijablu. Grupa autora je ovaj problem disbalansa opisala u svom radu [3]. Takav slučaj imamo i u podacima koji se koriste za ovaj eksperiment, gdje imamo omjer 88:12 neprelaznika u odnosu na prelaznike.
Ciljani atribut u setu podataka definiran je kao binarni tip podatka, pa su prelaznici označeni kao „true“, a neprelaznici kao „false“. Da bi nam modeli imali zadovoljavajuću točnost, potrebno je riješiti problem disbalansa u podacima.
U ovom eksperimentu je napravljeno poduzorkovanje (eng. *undersampling*) neprelaznika i to na način da je slučajnim izborom iz baze neprelaznika, izabrano 6750 zapisa.
Za učenje modela koristimo 70% zapisa, a 30% zapisa ostavljamo za testiranje. 70% zapisa za učenje sadrži 34908 zapisa. Od tog broja zapisa imamo 4430 zapisa označenih kao prelaznici. Svi prelaznici se koriste za eksperiment.
30478 zapisa je označeno kao neprelaznici i slučajnim poduzorkovanjem je izabrano 6750 zapisa. Na takav način, u setu podataka za učenje dobiven je omjer prelaznika i neprelaznika od 40:60. Cilj poduzorkovanja je povećanje točnosti predviđanja prelaznika.

B. Definicija churn-a i eksperiment

Definicija churn-a uključuje jedan od sljedeća tri navedena uvjeta:

a) Potrošnja na usluge SMS-a i Voice je manja od 1 KM,

b) broj dana aktivnosti je manji u odnosu na prosjek posljednja 3 mjeseca,

c) prosječan broj nadoplata zadnja tri mjeseca je manji od 5 KM i nije bilo nadoplata u mjesecu u kojem se definira churn.

Ulazni set podataka ima 109 različitih atributa. Zbog velikog broja atributa napravljena je selekcija korisnih atributa dok se ostali atributi ne koriste za izgradnju data mining modela. Selekcija atributa je napravljena korištenjem analize glavnih komponenti (eng. *Principal Component Analysis - PCA*). Analiza glavnih komponenti pronalazi smjer u kojem imamo najveću varijancu u visokodimenzionalnom setu podataka i projicira je na manji set podataka koji zadržava većinu bitnih informacija iz inicijalnog seta podataka. U ovom radu je napravljena selekcija svih atributa koji imaju težinski faktor veći od 0.03, pa se na takav način selekcije došlo do konačne brojke od 45 atributa za izgradnju Data mining modela, koji su prikazani u tabeli 2.

TABELA II. POPIS ATRIBUTA

POPIS ATRIBUTA (nakon selekcije)
Conversation_Duration_For_Outgoing_Calls_To_VAS
Duration_Of_Incoming_Calls_From_Eronet
Duration_Of_Incoming_Calls_From_FBHT
Duration_Of_Incoming_Calls_From_M064
Duration_Of_Incoming_Calls_From_MTS
Duration_Of_Outgoing_Calls_To_Eronet

Duration_Of_Outgoing_Calls_To_Eronet_CallCenter
Duration_Of_Outgoing_Calls_To_FBHT
Duration_Of_Outgoing_Calls_To_FHT
Duration_Of_Outgoing_Calls_To_FTS
Duration_Of_Outgoing_Calls_To_M064
Duration_Of_Outgoing_Calls_To_MBHT
Duration_Of_Outgoing_Calls_To_MTS
Duration_Of_Outgoing_Calls_To_Other_BH_Mobile_CallCenter
Number_Of_Data_Transfered_Units
Number_Of_Distinct_Incoming_Eronet_Users
Number_Of_Distinct_Incoming_M064_Users
Number_Of_Distinct_Incoming_MBHT_Users
Number_Of_Distinct_Incoming_MTS_Users
Number_Of_Distinct_Outgoing_Eronet_Users
Number_Of_Distinct_Outgoing_FBHT_Users
Number_Of_Distinct_Outgoing_FHT_Users
Number_Of_Distinct_Outgoing_FTS_Users
Number_Of_Distinct_Outgoing_M064_Users
Number_Of_Distinct_Outgoing_MBHT_Users
Number_Of_Distinct_Outgoing_MTS_Users
Number_Of_Distinct_Outgoing_Other_Mobile_Users
Number_Of_Incoming_Calls_From_Eronet
Number_Of_Incoming_Calls_From_M064
Number_Of_Incoming_Calls_From_MBHT
Number_Of_Incoming_Calls_From_MTS
Number_Of_Outgoing_Calls_To_Eronet
Number_Of_Outgoing_Calls_To_Eronet_CallCenter
Number_Of_Outgoing_Calls_To_FBHT
Number_Of_Outgoing_Calls_To_FHT
Number_Of_Outgoing_Calls_To_FTS
Number_Of_Outgoing_Calls_To_M064
Number_Of_Outgoing_Calls_To_MBHT
Number_Of_Outgoing_Calls_To_MTS
Number_Of_Outgoing_Calls_To_Other_Mobile
Number_Of_Sessions
Number_Of_SMS_To_Eronet
Number_Of_SMS_To_M064
Number_Of_SMS_To_MBHT
Number_Of_SMS_To_MTS

Ulazni set podataka sadrži atribute koji imaju značajno veće vrijednosti od drugih atributa, npr. atribut „količina prenesenih podataka“ ima veće vrijednosti od atributa „broj poziva“. Vrijednosti atributa ovise o njihovoj prirodi i o mjernim jedinicama. Atributi sa većim vrijednostima imaju veći utjecaj na izgradnju modela. Da bi napravili pravilnu ravnotežu utjecaja svih atributa, napravljena je normalizacija. Normalizacijom su se vrijednosti atributa reskalirale unutar određenog raspona brojeva. Za potrebe ovog rada napravljena je *Z* transformacija ili kako se još zove, statistička normalizacija. Formula statističke normalizacije prikazana je izrazom:

$$Z = (X - \mu) / s \quad (1)$$

Gdje je:

- X - vektor vrijednost atributa za koji se računa Z transformacija;
- μ - aritmetička sredina koja se oduzima od vektora X;
- s - standardna devijacija;

Z transformacija pretvara podatke u normalnu distribuciju sa srednjom vrijednošću 0 i varijancom 1. Na ovakav način su sve vrijednosti atributa svedene na istu skalu i na takav način je moguće napraviti kvalitetnu usporedbu među njima.
U eksperimentu je napravljena usporedba više algoritama strojnog učenja.
Svi algoritmi su trenirani i testirani nad poduzorkovanim setom podatka i sa reduciranim setom atributa.
Korišteni algoritmi su: stablo odlučivanja sa algoritmom C4.5, logistička regresija i neuronska mreža.

C. Rezultati eksperimenta

U radu su korištene standardne mjere da bi usporedili efikasnost različitih modela predikcije prelaska kupaca. Mjere koje su korištene u radu su: *točnos*t, *preciznost* i *senzitivnost*. Konfuzijskom matricom su predstavljeni rezultati eksperimenta, gdje su u obzir uzeta tri algoritma: *stablo odlučivanja*, *logistička regresija* i *neuronska mreža*. Napravljena je i grafička usporedba korištenih algoritama prikazom ROC krivulje.

U tabeli 3. predstavljena je konfuzijska matrica za model stabla odlučivanja:

TABELA III. STABLO ODLUČIVANJA C4.5

	FALSE	TRUE
FALSE	12403	316
TRUE	659	1582

Ukupna točnost modela stabla odlučivanja je 93,69%, senzitivnost je 83,40%, a preciznost je 71,56%.
Model stabla odlučivanja nam prikazuje pravila koja su korištena za podjelu podataka na homogenije skupove s obzirom na ciljanu varijablu churn. Analizom izgrađenog

stabla odlučivanja, moguće je zaključiti da su najvažniji atributi broj i trajanje poziva unutar mreže. Korišten je omjer dobiti (eng. *gain ratio)* kao kriterij na osnovu kojeg se biraju atributi za podjelu podataka.

U tabeli 4. predstavljena je konfuzijska matrica za model logističke regresije:

TABELA IV. LOGISTIČKA REGRESIJA

	FALSE	TRUE
FALSE	11962	439
TRUE	1100	1459

Logistička regresija je pokazala najlošije rezultate u predviđanju prelazaka korisnika. Ukupna točnost logističke regresije je 89,71%, senzitivnost je 76,87%, a preciznost je 57,01%.

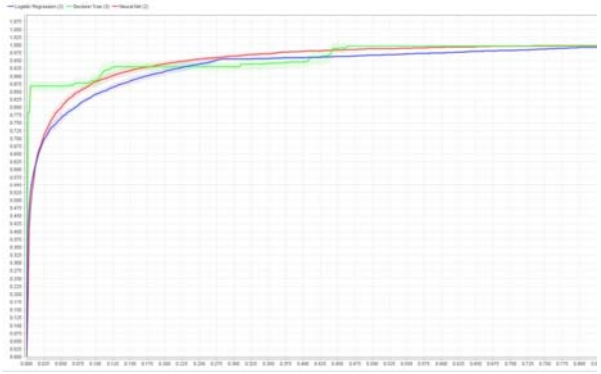
U tabeli 5. predstavljena je konfuzijska matrica za model neuronske mreže:

TABELA V. NEURONSKA MREŽA

	FALSE	TRUE
FALSE	12098	286
TRUE	964	1612

Izgrađeni model neuronske mreže ima jedan ulazni, jedan skriveni i jedan izlazni sloj. Ukupna točnost modela neuronske mreže je 91,64%. Neuronska mreža ima najveći vrijednost mjere senzitivnosti u vrijednosti od 84,93% tj. od svih testiranih modela ima najveći postotak predviđanja prelaznika. Neuronska mreža ima nešto lošiju mjeru preciznosti 62.58% u odnosu na stablo odlučivanja.

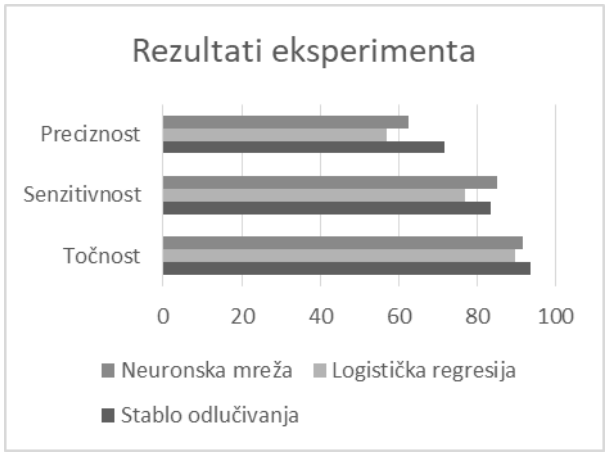
Na slici 1 je prikazana ROC (eng. *Receiver Operating Characteristics*) krivulja koja grafički prikazuje odnos senzitivnosti i specifičnosti za sva tri korištena algoritma strojnog učenja. Što je ROC krivulja algoritma bliže točki (0,1) to je algoritam bolji [11].



Slika 1. ROC krivulja usporedbe tri korištena algoritma strojnog učenja

Na ROC krivulji stablo odlučivanja je prikazano zelenom bojom, logistička regresija plavom i neuronska mreža crvenom bojom.

Na slici 2 prikazan je odnos između mjera točnosti, preciznosti i senzitivnosti sva tri korištena algoritma strojnog učenja.



Slika 2. Prikaz mjera efikasnosti tri korištena algoritma strojnog učenja

Rezultati su pokazali kako stablo odlučivanja daje najveći postotak točnosti i preciznosti, pa možemo zaključiti da se stablo odlučivanja pokazalo kao najbolji algoritam strojnog učenja za ovaj set podataka.

IV. ZAKLJUČAK

U radu je napravljena usporedba tri različita algoritma strojnog učenja da bi se predvidjeli prelasci prepaid telekom kupaca. Korišteni set podataka posjeduje disbalans u podacima gledajući ciljani atribut churn.

Problem disbalansa je riješen poduzorkovanjem i na takav način se došlo do omjera 40 % prelaznika u odnosu na 60 % neprelaznika. U radu je korištena statistička metoda analize glavnih komponenti koja je pomogla da se smanji broj atributa sa 109 na 45.

Napravljena je usporedba performansi modela koristeći mjere točnosti, senzitivnosti i preciznosti. Za svaki model je prikazana i konfuzijska matrica. Eksperiment je pokazao da najveću točnost i preciznost ima model stabla odlučivanja. Neuronska mreža ima neznatno veću mjeru senzitivnosti od stabla odlučivanja ali to ne mijenja činjenicu da se stablo odlučivanja pokazalo kao najbolji model za ovaj set podataka. Logistička regresija je pokazala najlošije rezultate u ovom eksperimentu.

Svaki korak u procesu predikcije prelaska korisnika je jako bitan, a poseban naglasak je na koracima redukcije atributa i rješavanja problema disbalansa u podacima.

V. BUDUĆE ISTRAŽIVANJE

Buduće istraživanje se može fokusirati na rješavanje problema disbalansa u podacima sa naduzorkovanjem i usporediti različite algoritme strojnog učenja sa poduzorkovanjem i naduzorkovanjem. Također je moguće

usporediti i dodatne algoritme strojnog učenja. Pošto su modeli za predikciju prelazaka kupaca jako ovisni o podacima, u budućem istraživanju bi bilo dobro koristiti dodatne setove podataka za analizu uspješnosti različitih modela.

LITERATURA

[1] V. Lazarov and M. Capota, "Churn Prediction," Eighth ACM SIGKDD Int. Conf., 2007.

[2] K. Dahiya, "Customer Churn Analysis in Telecom Industry," 2015.

[3] A. Amin et al., "Comparing oversampling techniques to handle the class imbalance problem□: A customer churn prediction case study," vol. 3536, no. c, pp. 1–18, 2016.

[4] H. Li, D. Yang, L. Yang, Y. Lu, and X. Lin, "Supervised Massive Data Analysis for Telecommunication Customer Churn Prediction," pp. 163–169, 2016.

[5] S. A. Qureshi, A. S. Rehman, A. M. Qamar, and A. Kamal, "Telecommunication Subscribers ' Churn Prediction Model Using Machine Learning," pp. 131–136, 2013.

[6] V. Umayaparvathi and K. Iyakutti, "Applications of Data Mining Techniques in Telecom Churn Prediction," Int. J. Comput. Appl., vol. 42, no. 20, pp. 5–9, 2012.

[7] C. Kirui, L. Hong, W. Cheruiyot, and H. Kirui, "Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining," vol. 10, no. 2, pp. 165–172, 2013.

[8] L. Peng and Y. Xiaoyang, "Telecom Customer Chum Prediction Based on Imbalanced Data Re-sampling Method," pp. 229–233, 2013.

[9] J. Zhang, J. Fu, C. Zhang, X. Ke, and Z. Hu, "Not Too Late to Identify Potential Churners□: Early Churn Prediction in Telecommunication Industry," pp. 194–199, 2016.

[10] A. Backiel, Y. Verbinnen, B. Baesens, and G. Claeskens, "Combining Local and Social Network Classifiers to Improve Churn Prediction," Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 2015 - ASONAM '15, pp. 651–658, 2015.

[11] D. DeLong, and D. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," International Biometric Society, 44(3):837–845, 1988.

ABSTRACT

Due to the high competition in the market, telecom operators are affected by the churn, so it is very important for them to identify which users are likely to leave and move to competitive telecoms. This research utilizes behavioral data from telecom systems that are used to identify patterns in behaviors and thereby recognize user transitions. When preparing data, a selection of useful attributes was made using the Principal Component Analysis (PCA). Also, the normalization of attribute values has been made to achieve a proper balance of the influence of all attributes. Several prediction models for detecting Churna Prepaid users in the telecom have been created in the paper and a performance analysis of the implemented Data mining models was performed.

DEVELOPMENT OF THE DATA MINING MODEL FOR CHURN PREDICTION OF PREPAID USERS IN THE TELECOM

Marin Mandić, Goran Kraljević, Ivan Boban