Predviđanje uspjeha bankarskog marketinga

Davorin Gradečak
Prirodoslovno-matematički fakultet
Sveučilište u Zagrebu
Zagreb, Hrvatska
grdavor@student.math.hr

Petra Jambriško
Prirodoslovno-matematički fakultet
Sveučilište u Zagrebu
Zagreb, Hrvatska
japetra@student.math.hr

Jakov Krunić
Prirodoslovno-matematički fakultet
Sveučilište u Zagrebu
Zagreb, Hrvatska
kjakov@student.math.hr

***Sažetak*—Projektni zadatak predstavljen ovim izvješćem bavi se predviđanjem odgovora korisnika na upit o ugovaranju oročenog depozita. Cilj projekta je otkriti najbolji algoritam strojnog učenja koji bankama može donijeti najveći profit u tzv. telemarketingu. Primjenjivali smo logističku regresiju, naivni Bayes, SVM i slučajne šume. Dajemo osvrt na druge pristupe te navodimo smjernice za mogući nastavak istraživanja.**

***Index Terms*—strojno učenje, logistička regresija, naivni Bayes, SVM, slučajne šume, XGBoost, banka, marketing, oročeni depozit**

I. Uvod

Jedan od najpoznatijih marketinga u današnje vrijeme je takozvani telemarketing. Telemarketing je marketinška metoda kojom se interakcija s korisnicima odvija na daljinu, najčešće putem telefona ili mobitela. Smatra se popularnom metodom zbog niskih troškova. Marketinške kampanje važan su dio bankarskih poslovanja. U situaciji koju ćemo mi istražiti, banke su direktno kontaktirale svoje korisnike s ciljem da korisnik ugovori oročeni depozit. Korisnici su mogli biti kontaktirani i više puta, ali naši podaci će se odnositi samo na posljednji kontakt. Ideja ovog projekta je da metodama strojnog učenja napravimo što bolji prediktor te da otkrijemo s kakvim korisnicima banka ima najveće šanse za ugovaranje oročenog depozita (korisnici su opisani raznim atributima). U tu svrhu, komentirat ćemo i isplativost ovakve marketinške kampanje banki.

II. Opis problema

Problem kojim se bavimo je predviđanje uspjeha bankarskog marketinga na temelju popisnih podataka. Podaci koje koristimo su nastali u marketinškim kampanjama koje su se bazirale na telefonskim pozivima. Jedan od ciljeva našeg istraživanja je odrediti utječu li pojedini atributi iz skupa podataka na varijablu y koju predviđamo (target). Drugim riječima, zanima nas koliki je utjecaj određenih karakteristika korisnika na njegovu odluku o tome hoće li ugovoriti oročeni depozit ili ne.

Podatke preuzimamo s web stranice UCI Machine Learning Repository (novija verzija podataka u [2]). Podaci se odnose na izravne marketinške kampanje portugalske bankarske institucije. Skup podataka koji ćemo koristiti je gotovo identičan onome koji je korišten u [3].

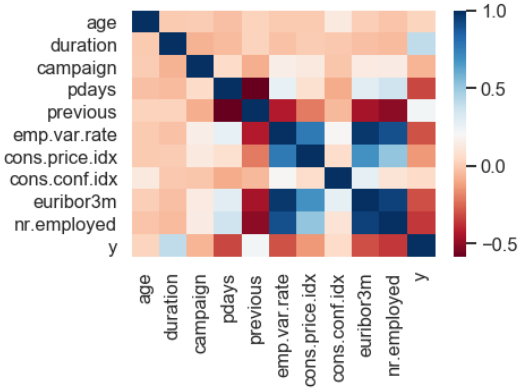
Podaci su podijeljeni u dva skupa podataka. To su *train* podaci i *test* podaci. Skup podataka *train* se sastoji od 41188 instanci, a skup podataka *test* se sastoji od 4119 instanci, tj. 10% instanci iz *train* skupa koje su odabrane nasumično. Svaka instanca predstavlja određenog korisnika banke. Svaki korisnik je opisan s 20 atributa i atributom y koji predviđamo (target). Atributi su sljedeći:

- age: dob korisnika, numerička varijabla
- job: vrsta posla korisnika, kategorijska varijabla (12 kategorija)
- marital: bračni status korisnika, kategorijska varijabla (4 kategorije)
- education: stupanj obrazovanja korisnika, kategorijska varijabla (8 kategorija)
- default: ima li korisnik neizvršene novčane obaveze, kategorijska varijabla (kategorije: 'no', 'yes', 'unknown')
- housing: ima li korisnik stambeni kredit, kategorijska varijabla (kategorije: 'no', 'yes', 'unknown')
- loan: ima li korisnik osobne zajmove, kategorijska varijabla (kategorije: 'no', 'yes', 'unknown')
- contact: tip komunikacije s korisnikom, kategorijska varijabla (kategorije: 'cellular', 'telephone')
- month: mjesec kad je obavljen zadnji kontakt s korisnikom, kategorijska varijabla (12 kategorija)
- day_of_week: dan u tjednu kad je obavljen zadnji kontakt s korisnikom, kategorijska varijabla (5 kategorija)
- duration: trajanje zadnjeg kontakta s korisnikom u sekndama, numerička varijabla
- campaign: broj kontaktiranja korisnika tijekom trenutne kampanje, numerička varijabla
- pdays: broj dana koji su prošli od zadnjeg kontaktiranja korisnika tijekom prethodne kampanje, numerička varijabla (999 znači da nisu bili kontaktirani prije)
- previous: broj kontaktiranja korisnika prije trenutne kampanje, numerička varijabla
- poutcome: ishod prethodne kampanje, kategorijska varijabla (kategorije: 'failure', 'nonexistent', 'success')
- emp.var.rate: stopa varijacije zaposlenosti-kvartalni indikator, numerička varijabla
- cons.price.idx: indeks potrošačkih cijena-mjesečni indikator, numerička varijabla
- cons.conf.idx: indeks povjerenja potrošača-mjesečni indikator, numerička varijabla
- euribor3m: euribor tromjesečna stopa-dnevni indikator, numerička varijabla

- nr.employed: broj zaposlenih-kvartalni indikator, numerička varijabla
- y: podatak koji predviđamo (**target**), je li korisnik ugovorio oročeni depozit ili ne, kategorijska varijabla (kategorije: 'yes', 'no')

Uočavamo su atributi vezani uz samog korisnika (godine, spol, vrsta posla ...), uz zadnji kontakt banke s korisnikom tijekom kampanje (dan u tjednu, trajanje ...), uz prethodne kontakte s korisnikom (broj kontakata prije i tijekom trenutne kampanje ...) i uz neke ekonomske podatke tijekom trenutne kampanje (indeks potrošačkih cijena ...).

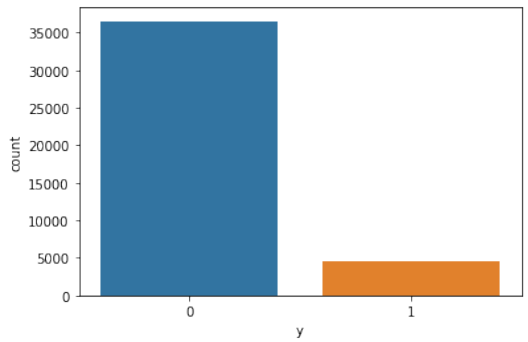
Numeričke vrijednosti možemo prikazati u korelacijskoj matrici. Prikazujemo koliko su korelirani određeni atributi da steknemo bolji dojam o podacima.



Slika 1: Korelacijska matrica

Najveću korelaciju s target varijablom ima atribut duration (trajanje razgovora s korisnikom). Također vidimo da su atributi vezani uz ekonomiju međusobno jako korelirani, što je bilo i za očekivati.

Prije samog prelaska rješavanja problema, važno je naglasiti da se radi o izrazito nebalansiranom skupu podataka. Bolji dojam o tome nam daje sljedeća slika.



Slika 2: Raspodjela pozitivnih i negativnih instanci

III. Metode i pristup rješavanju problema

Budući da je vrijednost koju naši algoritmi trebaju predvidjeti binarna varijabla, problem koji obrađujemo je klasifikacijski. Problem smo pokušali riješiti sljedećim algoritmima:

- Logistička regresija
- Naivni Bayes
- SVM
- Stabla odluke i slučajne šume

Uspješnost algoritama mjerili smo pomoću ROC-AUC score-a te pomoću AUPRC score-a. Budući da podaci nisu balansirani (89% ispitanih korisnika nije ugovorilo oročeni depozit, a 11% korisnika jest), kao glavnu mjeru evaluacije uzeli smo AUPRC (area under precision-recall curve) score. Prikazat ćemo i krivulju precision-recall za svaki algoritam. Preciznost (precision) je udio stvarno pozitivnih primjera u svima koji su modelom predviđeni kao pozitivni, a osjetljivost (recall) je udio točno pozitivnih primjera koje je model prepoznao kao pozitivne od ukupnog broja pozitivnih primjera.

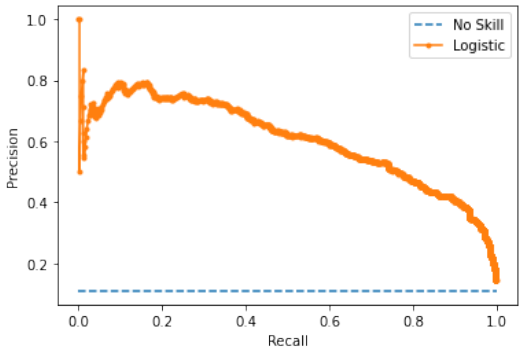
AUPRC je realan broj između 0 i 1. Što je veći AUPRC, to je naš model bolji. Važno je napomenuti da je ova metrika znatno drugačija od AUC-a. Ne postoji fiksna vrijednost za loš AUPRC score. U našoj situaciji, za trivijalni klasifikator koji za sve korisnike predviđa pozitivan odgovor AUPRC iznosi 0.11. Time možemo očekivati da AUPRC rezultati naših algoritama neće biti pretjerano veliki, no to ne znači da su naši prediktori loši jer već sve iznad 0.11 je bolje od trivijalnog.

Što se tiče samog skupa podataka, uočili smo da nema nedostajućih vrijednosti (missing values), no otkrili smo da ima dupliciranih podataka koje smo onda izbacili (12 redaka).

Da bismo mogli primjenjivati modele strojnog učenja na podatke, morali smo napraviti takozvani 'encoding' kategorijskih varijabli. U tu svrhu koristili smo One Hot Encoding.

A. *Logistička regresija*

Prvi algoritam s kojim smo pokušali jest logistička regresija. Najprije uočavamo da su preciznost i AUC score vrlo visoki (iznad 0.9). No ono što nas najviše zanima jest AUPRC vrijednost za koju smo dobili da iznosi 0.6 što možemo smatrati vrlo dobrim. Uz to prikazujemo i precision-recall krivulju.

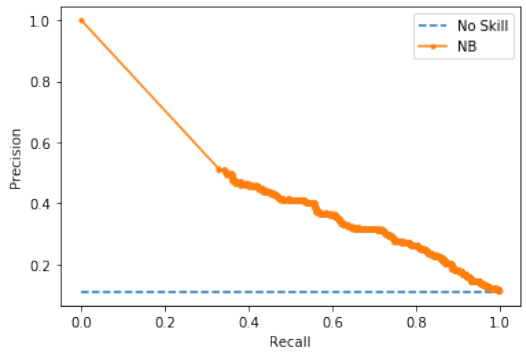


Slika 3: Precision-recall krivulja za logističku regresiju

B. *Naivni Bayes*

Kod naivnog Bayesa odmah uočavamo da su preciznost i AUC score lošiji nego kod logističke regresije (ispod 0.9) te

uočavamo velik broj false positive primjera. Konačnu potvrdu o tome da je naivni Bayes znatno lošiji dobivamo iz precision-recall krivulje i AUPRC vrijednosti koja iznosi 0.46.



Slika 4: Precision-recall krivulja za naivni Bayes

C. SVM

Kao i algoritam naivnog Bayesa, tako i SVM algoritam ne pokazuje bolje rezultate od logističke regresije. Osim toga, potrebno mu je i znatno duže vrijeme za učenje na danom skupu primjera. Za neuravnotežene skupove podataka, što mi imamo za slučaj, razdvajajuća hiperravnina dobivena pomoću SVM može biti pristrana prema predviđanjima većinske klase na testnim podacima. To je upravo razlog zašto dobivamo slabe rezultate (ako usporedimo s time da ako predviđamo samo negativne ishode, dobili bismo točnost od 0.8, a SVM algoritmom dobivamo 0.88).

	Metric	Score Linear Without PCA	Score RBF Without PCA	Score RBF With PCA (31)
0	Accuracy	0.881282	0.900947	0.914300
1	Precision	0.455814	0.630303	0.658065
2	Recall	0.434590	0.230599	0.452328
3	F1 Score	0.444949	0.337662	0.536137
4	AUPRC	0.448848	0.582164	0.586098

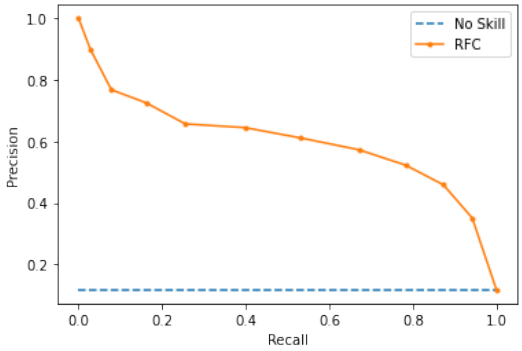
Slika 5: Rezultati dobiveni SVM algoritmima

D. Stabla odluke, slučajne šume i XGBoost

Kako naši podaci nisu uravnoteženi, potrebno je koristiti klasifikator koji je robusan prema disbalansu klasa. Najočitije je stablo odluka. Ako se rijetka klasa nalazi u određenom području obilježja ili barem to obično čini, većina ili sve rijetke klase leže u jednom čvoru stabla odlučivanja.

Na danom skupu primjera za testiranje ne bismo ništa otkrili o našem algoritmu (elementi skupa za testiranje su nasumično odabrani elementi skupa za treniranje te bi zato svaki element bio u potpunosti točno klasificiran) pa smo skup za treniranje *train_test_split* metodom podijelili na nove skupove za treniranje i testiranje gdje 20% primjera ode u novi *test*, a ostatak u novi *train* skup.

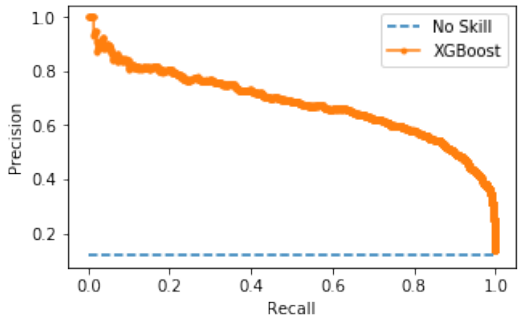
Ono što smo dobili alagoritmom stabla odlučivanja jest vrijednost AUPRC-a 0.56. Nadalje, slučajne šume uključuju odabir bootstrap uzoraka iz skupa podataka za treniranje i postavljanje stabla odluke za svaki takav uzorak. Ne koriste se sve značajke (varijable ili stupci), već umjesto toga, za svaki bootstrap uzorak odabran je mali nasumično odabrani podskup mogućnosti (stupaca). Kao posljedica, uklanja se korelacija stabala odluke (što ih čini neovisnijima), a zauzvrat, poboljšava se predviđanje ansambla. Ovdje je AUPRC ispao 0.6028, a na sljedećoj slici prikazujemo pripadnu precision-recall krivulju.



Slika 6: Precision-recall krivulja za slučajne šume

Ono što smo dalje pokušali jest probali dati strožu kaznu za pogrešno klasificiranje manjinske klase, što se naziva ponderiranom slučajnom šumom, no to, a niti ponderiranje klase na temelju distribucije klasa u svakom uzorku, nije dalo bolje rezultate. Isto tako, pokušali smo slučajno smanjivanje uzoraka većinske klase u bootstrap uzorku, ali ni to nije dalo bolju vrijednost AUPRC-a. Ono što jest bilo bolje, za zadnju metodu dodanu na slučajnu šumu jest srednja vrijednost ROC-AUC-a u odnosu na običnu slučajnu šumu, dobivena ponovljenom unakrsnom validacijom slojevitog K-folda.

Nadalje, XGBoost je kraći zapis za Extreme Gradient Boosting i to je ansambl stabala gdje nova stabla ispravljaju pogreške onih stabala koja su već dio modela. Stabla se dodaju sve dok se daljnja poboljšanja ne mogu napraviti. Za ovaj smo model dobili bolji rezultat, tj. AUPRC vrijednost je 0.667031. Uz to prikazujemo i precision-recall krivulju kako bi stekli još bolji dojam o uspješnosti algoritma.



Slika 7: Precision-recall krivulja za XGBoost

E. Promjena funkcije troška

Osim robusnijih klasifikatora, postoje i druge mogućnosti za poboljšanje klasifikatora. Možemo promijeniti ciljnu funkciju tako da uzmemo u obzir da trošak false pozitivnih i false negativnih vrijednosti ne može biti isti, a trošak oba može varirati ovisno o detaljima iz primjera. U našem slučaju, to bismo mogli objasniti ovako: cijena false pozitivnog može biti trošak vremena za pozivanje takvih korisnika, dok trošak false negativnog može biti gubitak korisnika koji bi ugovorio depozit, a zbog procjene ga se ne nazove.

Možemo promijeniti svoju ciljnu funkciju da to odražavamo. Na primjer, standardna funkcija gubitka logističke regresije je negativna log-likelihood funkcija

$$\sum_{i=1}^{N}[-y_i \log(h(X)^{(i)})) - (1-y_i)\log(1-h(X)^{(i)})),]$$

gdje $h(X)^{(i)}$ označava $P(y=1|X^{(i)})$.

Međutim, ako kažemo da svako opažanje ima neki true pozitivan C_{TP_i}, true negativan C_{TN_i}, false pozitivan C_{FP_i} i false negativan trošak C_{FN_i} povezan s njim, možemo koristiti sljedeću objektivnu funkciju:

$$\frac{1}{N}\sum_{i=1}^{N}[y_i\left(\log(h(X)^{(i)})C_{TP_i} + (1-\log(h(X)^{(i)})C_{FN_i})\right)$$
$$+(1-y_i)\left(\log(h(X)^{(i)})C_{FP_i} + (1-\log(h(X)^{(i)})C_{TN_i})\right)].$$

Za trošak true positivea i true negativea smo stavili 0. Za trošak false negativea smo stavili fiksnu vrijednost od 1000 (eura) koji predstavlja gubitak profita zbog propusta klijenta. Trošak false positivea smo postavili kao okvirni trošak poziva prema tom potencijalnom klijentu. Konkretnije, pretpostavili smo da je cijena poziva 30 eura po satu, plaća teleoperatera 4.38 eura po satu, to pomnožili sa umnoškom ukupnog broja poziva tom klijentu i trajanju zadnjeg poziva. Očito je da su ove ocijene dosta primitivne. Navedimo par razloga:

- Za trošak false positivea smo pretpostavili da bi svaki klijent dao isti depozit, ili da bi dao prosjek depozita. To definitivno ne mora vrijediti te bi mogli modelirati količinu depozita prema nekim ekonomskim featurima. Nažalost, u ovom datasetu nema konkretnih ekonomskih featura koji su direktno vezani za mogućeg klijenta, već samo po vremenskom periodu kada je kontaktiran. Nažalost, iz featurea koje imamo nije očito kako bi modelirali količinu depozita.
- Kod cijene poziva smo pretpostavili da je svaki poziv nekog klijenta trajao jednako dugo kao i zadnji, što naravno nije točno. Ipak, ako ne želimo pretpostaviti da svaki razgovor traje jednako dugo bez obzira na klijenta, ovo se čini kao prihvatljiv kompromis.
- Kod računanja cijene false negativea bi mogli uračunati vjerojatnost da potencijalnog klijenta izgubimo zauvijek, ako nastavimo zvati taj broj

Usprkos ovim manama, smatramo da omjer cijene između false positive i false negative relativno dobro reflektira stvarno stanje, tj. da je gubitak mogućeg klijenta za nekoliko reda veličine skuplji od cijena poziva.

Tako definiran trošak smo isprobali na modelu stabla odluke i običnih slučajnih šuma te dobili sljedeće rezultate, tj. razlike u cijenama na test podacima: običnim stablom odluke s gore navedenim cijenama dobili bismo trošak 466392.99 jedinica, a običnom slučajnom šumom 584525.402 jedinica. Nadalje, ovim troškovnim učenjem, cijena stabla odluke i slučajne šume iznosi 36512.73465, odnosno 35452.0257 jedinica što je poprilično poboljšanje.

Osim na ovim modelima, odlučili smo opisanu metodu isprobati i na logističkoj regresiji gdje ponovo dobivamo bolji rezultat, tj. manji trošak na skupu za testiranje.

IV. Rezultati

U ovom odjeljku sumirat ćemo sve dosad dobivene rezultate. Za sve algoritme koje smo proveli navodimo točnost i AUPRC vrijednost u sljedećoj tablici.

	Accuracy	AUPRC
Logistička regresija	0.913086	0.601339
Naivni Bayes	0.854819	0.463814
SVM Linear	0.881282	0.448848
SVM RBF	0.900947	0.582164
SVM RBF (PCA)	0.914300	0.586098
Slučajne šume	0.903594	0.601897
Bootstrap slučajne šume	0.903473	0.611274
XGBoost	0.911789	0.667031

Tablica I: Usporedba rezultata korištenih metoda

S obzirom da smo već rekli da je AUPRC metrika po kojoj ćemo mjeriti uspješnost naših algoritama, iz prethodne tablice možemo zaključiti da je najuspješniji algoritam XGBoost s AUPRC rezultatom 0.667031. Uočavamo i da XGBoost nije najbolji što se tiče točnosti (accuracy). To nije čudno jer znamo da velika točnost ne znači nužno veliku AUPRC vrijednost. Najlošijim algoritmom smatramo naivni Bayes jer smo SVM ipak uspjeli poboljšati uz neke izmjene. Ipak, nedostatak SVM-a je taj što mu je potrebno znatno duže vremena za učenje stoga možemo reći da on nije dovoljno efikasan.

V. Osvrt na druge pristupe

Naš pristup je sličan pristupu u radu [3]. U tom radu naglasak je na feature engineeringu. Ispitane su četiri klasifikacijske metode. To su logistička regresija, stablo odlučivanja, neuronske mreže i metoda potpornih vektora. Ove metode su uspoređene pomoću dvije klasifikacijske metrike: AUC (Area Under Curve) i ALIFT (Area of the LIFT cumulative curve). Za obje metrike, najbolji rezultati su ostvareni metodom neuronskih mreža (AUC = 0.8 i ALIFT = 0.67). To je jedna od razlika u odnosu na naš pristup budući da je naša glavna klasifikacijska metrika AUPRC (Area Under Precision-Recall Curve). Također, u tom radu je korišten drugačiji skup podataka s više instanci i više atributa, tj. korisnici i njihova ekonomska pozadina su detaljnije opisani. Stoga ne možemo

jasno reći jesu li naši rezultati bolji ili ne. Naglašavamo da se i u radu [3] također koristi nebalansirani skup podataka (12.38% pozitivnih instanci).

Rad [1] se bazira na usporedbi više različitih algoritama: algoritam slučajnih šuma, metoda potpornih vektora, neuronske mreže, naivni Bayesov algoritam i algoritam k-najbližih susjeda. Rezultati ovog rada su pokazali da se najbolja predviđanja postižu metodom potpornih vektora i naivnim Bayesovim algoritmom. Zbog jednostavnosti i održivosti, preferira se naivni Bayesov algoritam. Korišten je isti skup podataka kao i naš. Naši rezultati su ipak znatno drugačiji od rezultata ovog rada budući da smo metodu potpornih vektora i naivni Bayesov algoritam svrstali među lošije metode.

Jedan od popularnijih pristupa koje smo pronašli je korištenje metode XGBoost (Extreme Gradient Boosting Decision Tree) koja je nama dala najbolje rezultate. Korištena je metrika AUC, a rezultati su se pokazali vrlo dobrima (AUC > 0.8). Stoga smatramo da su i naši rezultati vrlo uspješni jer je upravo ova metoda najbolja u našem projektu. Ipak, za još bolju usporedbu, nedostaju AUPRC vrijednosti koje nisu korištene u drugim radovima.

Zanimljivo je da se u radovima koje smo pronašli ne računa profit. Ne spominju se nikakve cijene niti troškovi kampanje, a toga smo se mi ipak dotakli. U drugim radovima je naglasak najviše bio na usporedbi algoritama te na pronalasku najboljeg algoritma. Zaključujemo da je u svim radovima naglasak ipak na teoriji, dok je praktična realizacija nažalost nedovoljno istražena. Međutim, za praktičnu realizaciju je nužna uska povezanost s nekom bankarskom institucijom što još nije ostvareno.

VI. Mogući nastavak istraživanja

Telemarketing je jedna od popularnih metoda ne samo u bankarskom sektoru, već i u raznim drugim područjima. Cilj je svima isti—maksimizirati profit. U današnje vrijeme, olakotna okolnost je to što su troškovi telemarketinga jako niski. Cijene telefonskih poziva su praktički zanemarive, a stvarni trošak je samo vrijeme koje banka i njeni agenti mogu izgubiti. Time svaki pozitivan odgovor klijenta dobiva još veću vrijednost. Osim u ovom našem slučaju, prediktori bi se mogli primijeniti u drugim oblicima marketinga i u situacijama gdje se radi o velikim troškovima.

U našem istraživanju nismo imali podatke o cijenama telefonskih poziva tj. troškova i nismo imali podatke o mogućim profitima za svaki pozitivan odgovor. Zbog toga nismo mogli napraviti preciznu analizu dobiti i isplativosti koja bi se mogla prezentirati bankarima. Jedan od mogućih nastavaka je napraviti univerzalni model koji će moći primijenjivati svaka banka.

Za bolje prediktore smatramo da bi trebalo biti mnogo više atributa. Osim toga, smatramo da bi se trebali proučavati i agenti tj. ljudi iz banke koji kontaktiraju korisnike. Korisno bi bilo proučiti njihove karakteristike (spol, dob, itd.) te istražiti koliki oni imaju utjecaj na odgovor korisnika. Još bi se mogla proučavati uspješnost pojedinih agenata te veza između agenata i njihovih uspješno kontaktiranih korisnika.

Na kraju zaključujemo da je potrebno uzeti još mnogo raznih faktora u obzir kako bi se ovakvi modeli zaista mogli primijeniti u praksi. Naglašavamo i da je važno detaljnije povezivanje s određenom bankarskom institucijom kako bi se mogao steći stvarni dojam o isplativosti ovakvih algoritama i prezentirati realno rješenje.

Literatura

[1] A. Abu-Srhan, S. Al zghoul, B. Alhammad, i R. Al-Sayyed. (2019). *Visualization and Analysis in Bank Direct Marketing Prediction*. Dostupno na https://thesai.org/Downloads/Volume10No7/Paper_85-Visualization_and_Analysis_in_Bank_Direct_Marketing.pdf

[2] http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[3] S. Moro, P. Cortez i P. Rita. (2014). *A data-driven approach to predict the success of bank telemarketing*. Dostupno na http://media.salford-systems.com/video/tutorial/2015/targeted_marketing.pdf

[4] https://github.com/pmf-strojnoucenje/Vjezbe

[5] https://glassboxmedicine.com/2019/03/02/measuring-performance-auprc/

[6] https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

[7] https://github.com/albahnsen/CostSensitiveClassification