Sveučilište u Zagrebu

Prirodoslovno matematički fakultet

Diplomski studij Financijska i poslovna matematika

Logistička regresija za binarne ishode

Brezak Klaudija, Briška Marija, Kao Tena Albina,

Labaš Maja, Mašović Julijana, Preksavec Jurica

Zagreb, lipanj 2023.

Sadržaj

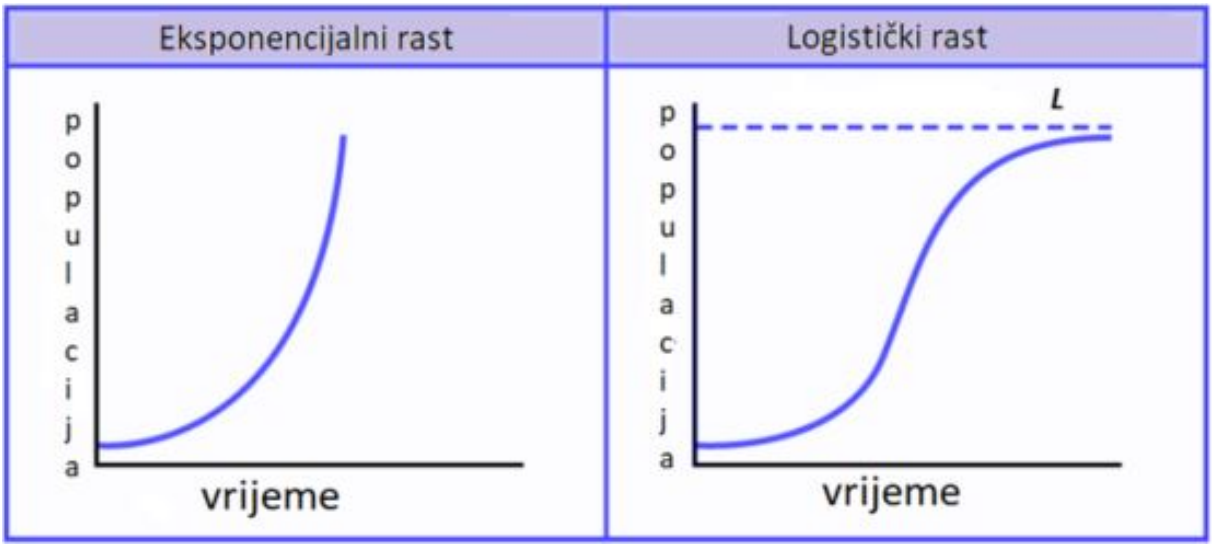
1 Uvod

1.1 Povijest logističke regresije

Logističku funkciju prvi je predstavio Belgijski matematičar Pierre Francois Verhulst sredinom 19. stoljeća. Predložio ju je kroz seriju od 3 znanstvena rada te je inicijalno služila za modeliranje rasta populacije, bilo ljudi, životinja ili biljaka. Verhulstova formula za logističku funkciju dana je s:

$$y = \frac{L}{1 + e^{-k(x-x_0)}}, \tag{1}$$

gdje je e baza prirodnog algoritma, x_0 je srednja vrijednost x-a (točka pregiba), L je maksimalna vrijednost od y, a k je maksimalni gradijent krivulje. U vrijeme kada je ova funkcija prvi puta spomenuta, smatralo se da eksponencijalni rast populacije mora jednom doći do zasićenja, odnosno da njezin eksponencijalni rast ne može vječno trajati, nego će se nakon određenog vremena značajno usporiti. što je prikazano na slici dolje:



Slika 1: Usporedba eksponencijalnog i logističkog rasta

Dolaskom 20. stoljeća, logistička funkcija pronašla je svoju primjenu u ekonomiji i kemiji, te u raznim drugim znanostima za modeliranje prirodnih pojava. Kao što se vidi na gornjoj slici s desne strane, krivulja logističke funkcije ima sigmoidalni oblik sličan kumulativnoj normalnoj distribuciji vjerojatnosti. Logistička regresija je statistički model

binarnog klasifikatora te se koristi za predviđanje ishoda dva moguća stanja (matematički ishod 0 ili 1, logički ishod da ili ne), te izlaz iz modela logističke regresije prikazujemo pomoću vjerojatnosti. Iz tog je razloga ovaj model vrlo brzo postao popularan kod modeliranja vjerojatnosnih pojava. Model logističke regresije temelji se, te se uvelike oslanja na podatkovne skupove na temelju kojih računamo statističku vjerojatnost.

2 Logistička regresija - definicija i interpretacija

Logička regresija je vrsta statističke analize koja se koristi za predviđanje binarnih ishoda ovisne varijable, kao što su da ili ne, na temelju prethodnih opažanja. Na primjer, algoritam može odrediti pobjednike olimpijskih igara na temelju analize rezultata prethodnih olimpijskih igara. Model logističke regresije može uzeti u obzir više kriterija. U slučaju pobjede u nekom sportu olimpijskih igara, logistička funkcija mogla bi uzeti u obzir čimbenike kao što su prethodne pobjede na natjecanjima ili koliko godina se osoba bavi treniranjem. Budući da se radi o predviđanju, vrijednost ovisne varijable je vrijednost iz segmenta [0,1]. U logističkoj regresiji se primjenjuje logit transformacija odnosno dijeli se vjerojatnost uspjeha sa vjerojatnosti neuspjeha. Logit transformacija daje linearnu relaciju između vjerojatnosti promatranog događaja p_i i vrijednosti nezavisne varijable X_i te se omjer $p_i/(1-p_i)$ naziva omjer šansi (odds ratio). Ova logistička funkcija predstavljena je kao:

$$logit(p_i) = \frac{1}{1+e^{-p_i}} \tag{2}$$

$$ln(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_k \cdot X_k \tag{3}$$

Logistička regresija postala je važan alat za strojno učenje na način da algoritmima koji se koriste u aplikacijama omogućuje klasificiranje dolaznih podataka na temelju povijesnih podataka. Algoritmi postaju bolji u predviđanju klasifikacija dolaskom dodatnih relevantnih podataka.

Popularna metoda za procjenu prikladnosti modela je Hosmer-Lemeshow test. Hosmer-Lemenshow test (HL test) je test za ocjenu prilagodbe za logističke regresije, posebno za modele predviđanja rizika. HL test govori koliko dobro podaci odgovaraju modelu. HL test koristi se samo za varijable binarnog ishoda.

OR predstavlja mogućnosti da će se ishod dogoditi, s obzirom na određeni događaj, u usporedbi s mogu˘cnostima da će se ishod dogoditi u odsutnosti tog događaja. Ako je OR veći od 1, tada je događaj povezan s većim mogu˘cnostima za generiranje određenog ishoda, a ukoliko je OR manji od 1, tada je događaj povezan s manjim mogu˘cnostima da će se taj ishod dogoditi.

3 Linearna vs. logistička regresija

I linearna i logistička regresija popularne su među algoritmima strojnog učenja te obje spadaju u istu vrstu učenja - učenje s nadzorom. Unatoč tome, postoje značajne razlike između ova dva modela. Linearna regresija određuje kakav je odnos između neprekidne zavisne varijable i jedne ili više nezavisnih varijabli. Ako se u modelu koristi samo jedna nezavisna varijabla govorimo o jednostavnoj linearnoj regresiji dok je kod dvije ili više nezavisnih varijabli riječ o višestrukoj linearnoj regresiji. Cilj linearne regresije je pronaći pravac koji najtočnije aproksimira dani skup podataka. Za to se najčešće koristi metoda najmanjih kvadrata. Pronalaskom pravca algoritam uspostavlja vezu između zavisne i nezavisnih varijabli koja bi trebala biti linearna. Logistička regresija također služi za određivanje odnosa između zavisne i nezavisnih varijabli, no ovdje je zavisna varijabla kategoričkog tipa. Tako na primjer zavisna varijabla može biti 0 ili 1, da ili ne, točno ili netočno i tako dalje. Za razliku od linearne regresije, kod koje ishod može poprimiti bilo koju od beskonačno mnogo vrijednosti, zavisne varijable i ishodi logističke regresije mogu poprimiti samo dvije vrijednosti (binarna logistička regresija). Logistička regresija nam daje vrijednost između 0 i 1, to jest daje vjerojatnost upadanja u neku od kategorija zavisne varijable. To se može prikazati S - krivuljom. Uz to, parametri ovog modela obično se procjenjuju metodom maksimalne vjerodostojnosti, a ne ranije spomenutom metodom najmanjih kvadrata. Nadalje, ovaj model ne zahtjeva linearan odnos među zavisnim i nezavisnim varijablama te nije dozvoljena kolinearnost nezavisnih varijabli dok linearna regresija to dopušta. Uz to, logistička regresija se mora provoditi na velikim uzorcima kako bi se mogao uočiti značajan učinak dok linearna regresija ne treba toliku veličinu uzorka.

Kako bi bolje razumjeli ključne razlike između linearne i logističke regresije ilustrirat

ćemo to na jednom primjeru. Pretpostavimo da promatramo rezultate studenta na nekom ispitu u odnosu na vrijeme koje je proveo učeći za ispit. U tom slučaju, linearna i logistička regresija mogu dati dvije različite vrste rezultata. S obzirom da linearna regresija daje kontinuirana predviđanja, pomoću nje možemo, ako nam je dana utrošena količina vremena na učenje, predvidjeti studentov rezultat na ispitu u obliku bodova (na primjer u rasponu od 0 do 100). S druge strane, budući da logistička regresija u svojim predviđanjima dopušta samo određene kategorije, ona bi nam kao rezultat mogla dati samo odgovor da li će student proći ili pasti ispit s obzirom na njegovo uloženo vrijeme. Na kraju možemo reći da se linearna regresija koristi kako bi se predvidjela neka kontinuirana, neprekidna vrijednost, a logistička regresija predviđa diskretnu vrijednost. Linearna regresija se odabire u slučaju rješavanja problema regresije, dok se logistička zapravo koristi za rješavanje problema klasifikacije.

4 Tipovi logističke regresije

Postoje dvije glavne vrste logističke regresije: binarna logistička regresija i multinomna logistička regresija. One se razlikuju u broju kategorija zavisne varijable.

4.1 Binarna logistička regresija

Kod binarne logističke regresije zavisna varijabla ima samo dva moguća ishoda, odnosno postoje samo dvije kategorije. Ovaj model je najčešće korištena vrsta logističke regresije te se koristi u mnogim domenama i sektorima. Na primjer, može se koristiti u marketinškoj analizi za prepoznavanje potencijalnih kupaca, u upravljanju ljudskim resursima za prepoznavanje zaposlenika za koje je moguće da će napustiti tvrtku, u upravljanju rizicima za predviđanje neplaćanja kredita, u osiguranju i tako dalje. Navedeni primjeri mogu se temeljiti na informacijama kao što su dob, spol, zanimanje, iznos premije, učestalost kupnje i slično. Neke od tih varijabli su kategoričke, a neke kontinuirane, ali u svim slučajevima zavisna varijabla je binarna.

4.2 Multinomna logistička regresija

U multinomnoj logističkoj regresiji zavisna varijabla ima tri ili više mogućih ishoda. Ovisno o kakvim se ishodima radi, odnosno ovisno kakve vrijednosti zavisna varijabla poprima, razlikujemo dvije vrste multinomne logističke regresije: logistička regresija s nominalnom zavisnom varijablom i logistička regresija s ordinalnom zavisnom varijablom. Nominalna zavisna varijabla poprima vrijednosti za koje ne postoji redoslijed kvalitete što bi značilo da nijedna kategorija nije više ili manje vrijedna od neke druge. Neki od primjera bi bili predviđanje koji smjer će budući gimnazijalac upisati (npr. jezična, prirodoslovno - matematička ili opća gimnazija, uzimajući u obzir njegove ocjene i najdraži predmet) ili predviđanje koju boju će osoba odabrati kao najdražu među tri ili više ponuđenih (npr. s obzirom na spol i dob osobe). Jasno je da ne postoji bolji i lošiji smjer niti bolja i lošija boja pa je za navedene probleme zaista prigodno koristiti nominalnu logističku regresiju. Ordinalna logistička regresija, isto kao i nominalna, podrazumijeva tri ili više kategorija kojima pripadaju vrijednosti zavisne varijable, ali u ovom slučaju te kategorije imaju točno definiran slijed. Na primjer, možemo se baviti predviđanjem tko će osvojiti brončanu, tko srebrnu, a tko zlatnu medalju na olimpijskim igrama (npr. s obzirom na godine treniranja i na broj do tada osvojenih medalja na drugim natjecanjima), predviđanjem rezultata ispita ocjenama od 1 do 5 (npr. s obzirom na broj izostanaka s predavanja i broj provedenih sati učeći) i slično. Budući da se više vrednuje ako sportaš osvoji zlato, a ne broncu ili ako učenik dobije peticu, a ne trojku jasno je da će najprimjereniji model za promatranje ovih problema uistinu biti ordinalna logistička regresija.

5 Logistička regresija i strojno učenje

U okviru strojnog učenja, logistička regresija pripada obitelji nadziranih modela strojnog učenja. Također se smatra diskriminirajućim modelom, što znači da pokušava razlikovati klase (ili kategorije). Za razliku od generativnog algoritma, kao što je naive bayes, on ne može, kao što ime implicira, generirati informacije, kao što je slika, klase koju pokušava predvidjeti (npr. slika mačke). Ranije smo spomenuli kako logistička regresija maksi-

mizira logaritamsku funkciju vjerojatnosti za određivanje beta koeficijenata modela. To se malo mijenja u kontekstu strojnog učenja. Unutar strojnog učenja, negativna logaritamska vjerojatnost koja se koristi kao funkcija gubitka, koristeći proces gradijentnog spuštanja da bi se pronašao globalni maksimum. Ovo je samo još jedan način da se dođe do istih procjena o kojima je gore bilo riječi. Logistička regresija također može biti sklona pretjeranom prilagođavanju, osobito kada postoji velik broj prediktorskih varijabli unutar modela. Regularizacija se obično koristi za kažnjavanje velikih koeficijenata parametara kada model pati od velike dimenzionalnosti.

6 Slučajevi korištenja logističke regresije

Logistička regresija se obično koristi za probleme predviđanja i klasifikacije. Neki od ovih slučajeva upotrebe uključuju:

- Otkrivanje prijevare: Logistički regresijski modeli mogu pomoći timovima da identificiraju anomalije podataka koje predviđaju prijevaru. Određena ponašanja ili karakteristike mogu imati veću povezanost s prijevarnim aktivnostima, što je osobito korisno bankarskim i drugim financijskim institucijama u zaštiti njihovih klijenata. Tvrtke koje se temelje na SaaS-u također su počele usvajati ove prakse kako bi eliminirale lažne korisničke račune iz svojih skupova podataka prilikom provođenja analize podataka o poslovnoj izvedbi.

- Predviđanje bolesti: U medicini se ovaj analitički pristup može koristiti za predviđanje vjerojatnosti bolesti ili bolesti za određenu populaciju. Zdravstvene organizacije mogu uspostaviti preventivnu skrb za pojedince koji pokazuju veću sklonost određenim bolestima.

- Predviđanje odljeva: Specifična ponašanja mogu biti pokazatelj odljeva u različitim funkcijama organizacije. Na primjer, timovi za ljudske resurse i menadžment možda bi željeli znati postoje li unutar tvrtke zaposlenici s visokim učinkom koji su u opasnosti da napuste organizaciju; ova vrsta uvida može potaknuti razgovore za razumijevanje problematičnih područja unutar tvrtke, kao što su kultura ili naknada.

Alternativno, prodajna organizacija možda će htjeti saznati koji su njihovi klijenti u opasnosti da svoje poslovanje prebace negdje drugdje. To može potaknuti timove da postave strategiju zadržavanja kako bi izbjegli gubitak prihoda.

7 Primjeri uspjeha logističke regresije

1. Procjena kreditnog rizika

   - Ako ste kreditni službenik u banci, tada želite biti u mogućnosti identificirati karakteristike koje su indikativne za ljude koji će vjerojatno kasniti s plaćanjem kredita i koristiti te karakteristike za prepoznavanje dobrih i loših kreditnih rizika.

2. Povećanje profita u bankarskoj industriji

   - First Tennessee Bank povećala je profitabilnost s IBM SPSS softverom i postigla povećanja do 600 posto u kampanjama unakrsne prodaje. Čelnici ove regionalne banke u SAD-u htjeli su pristupiti pravim klijentima s pravim proizvodima i uslugama. Nema manjka podataka za pomoć, ali bio je izazov premostiti jaz od posjedovanja podataka do poduzimanja radnji. First Tennessee koristi predistivnu analitiku i tehnike logističke analitike unutar analitičkog rješenja kako bi dobio bolji uvid u sve svoje podatke. Kao rezultat toga, donošenje odluka je poboljšano kako bi se optimizirala interakcija s klijentima.

8 Primjer I.

Određenom broju kupaca ponuđene su sezonske karte za zabavni park. Karte su se nudile u obliku promocijskog paketa koji uključuje besplatan parking te običnih sezonskih karata. Cilj je istražiti jesu li kupci više skloni kupovati sezonsku kartu kada je u određenom promocijskom paketu ili ne.
Kako bi postigli ikakav zaključak, potrebno je modelirati ovaj problem uz pomoć logističke regresije. Podatci se sastoje od načina obavljanja ponude (putem pošte, E-maila ili uživo u parku) te informacije je li ponuda sadržavala promocijski paket ili ne. Za početak ćemo

prikazati tablicu koja prikazuje uspješnost ponude na 3156 kupaca. Na prvoj slici možemo vidjeti kako je raspoređeno 1589 kupaca koji su prihvatili ponudu:

Nacin	Sa_paketom	Bez_paketa
Posta	242	359
Uzivo_u_parku	639	284
Email	38	27

Slika 2: Kupci koji su prihvatili ponudu

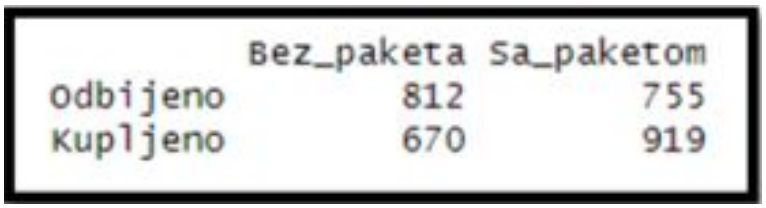
Na sljedećoj slici je prikazano 1567 kupaca koji su odbili ponudu:

Nacin	Sa_paketom	Bez_paketa
Posta	449	278
Uzivo_u_parku	223	49
Email	83	485

Slika 3: Kupci koji su odbili ponudu

8.1 Model I.

U ovom modelu nam način ponude nije bitan, stoga gledamo samo prisutnost promocijskog pameta te na sljedećoj slici možemo vidjeti broj prihvaćenih i odbijenih ponuda po dobivenoj opciji:



	Bez_paketa	Sa_paketom
Odbijeno	812	755
Kupljeno	670	919

Slika 4: Prikaz prihvaćenih i odbijenih ponuda

Kako bi prilagodili naše podatke za logističku regresiju, u R-u koristimo funkciju glm(). Zbog toga što je naš rezultat binaran, kao argument u funkciji glm() je potrebno koristiti "family=binomial". U ovom istraživanju zanima nas jesu li kupci skloniji uzeti sezonske

karte u promo paketu pa zbog toga uzimamo argumente "Pass" i "Promo" te na sljedećoj slici možemo vidjeti rezultate:

```
Call:
glm(formula = as.factor(Pass) ~ Promo, family = binomial, data = pass.df)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.19222    0.05219  -3.683 0.000231 ***
PromoSa_paketom  0.38879    0.07167   5.425 5.81e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4375.0  on 3155  degrees of freedom
Residual deviance: 4345.4  on 3154  degrees of freedom
AIC: 4349.4

Number of Fisher Scoring iterations: 3
```

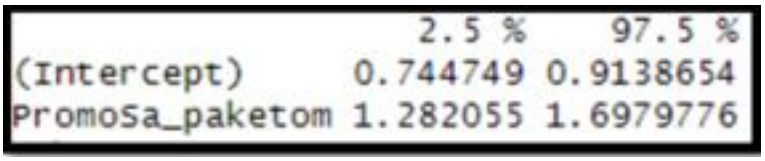
Slika 5: Rezultati glm() funkcije

Ono na što je potrebno obratiti pažnju su vrijednosti koeficijenata. Možemo vidjeti kako je vrijednost uz varijablu PromoSa_paketom pozitivna i iznosi 0.38879 što nam govori da je ostvaren pozitivan efekt kada je ponuda sezonskih karata uključena u neku vrstu paketa. Pomoću dobivenog koeficijenta možemo izračunati povezanost prodaje sezonskih karata i varijable promocijskog paketa u obliku omjera uspjeha i neuspjeha koristeći funkciju *plogis()*:

```
> plogis(0.3888) / (1-plogis(0.3888))
[1] 1.475209
```

Slika 6: Omjer uspjeha i neuspjeha

Dobiveni rezultat nam govori kako su kupci skloniji 1.475 puta (47.5%) više kupiti sezonske karte kada su uključene u promocijski paket.
Osim toga, možemo promatrati i pouzdane intervale za koeficijente:



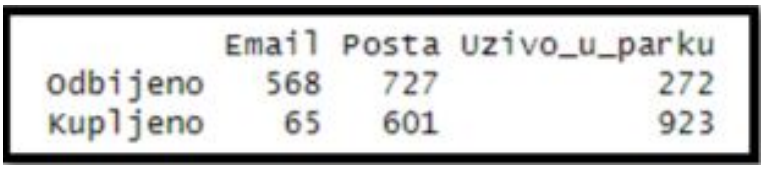
	2.5 %	97.5 %
(Intercept)	0.744749	0.9138654
PromoSa_paketom	1.282055	1.6979776

Slika 7: Intervali koeficijenta

Iz slike možemo vidjeti kako je interval za promocijski paket između 1.28 i 1.7 što nam govori kako postoji značajan pozitivan efekt. Problem nastaje kod činjenice da ne možemo biti sigurni što točno utječe na prodaju jer smo uzeli za pretpostavku da samo promocijski paket utječe što ne mora biti istina. S obzirom da postoji mogućnost da druge varijable mogu utjecati na rezultat, potrebno je provesti daljnje istraživanje.

8.2 Model II.

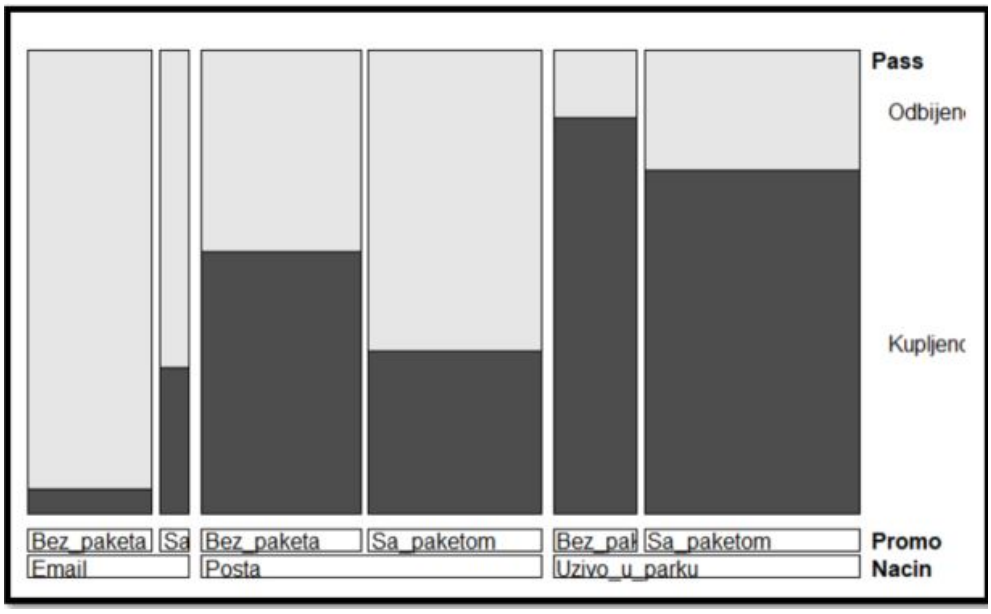
U ovom modelu uzimamo u obzir način ponude. Na sljedećoj slici su prikazane prihvaćene i odbijene ponude:



	Email	Posta	Uzivo_u_parku
Odbijeno	568	727	272
Kupljeno	65	601	923

Slika 8: Ponude po načinu

Možemo primijetiti kako je najuspješniji način ponude uživo u parku gdje je prihvaćeno čak 923 ponude. Drugi način na koji ove podatke možemo prikazati je pomoću double-decker plota za kojeg je potrebno instalirati paket vcdExtra:



Slika 9: Doubledecker plot

Na slici su prikazane uspješne (crna boja) i neuspješne (siva boja) prodaje. Također možemo primijetiti kako svaki način pristupanja ima različiti efekt na prodaju. Tako kod Email ponude se vidi da se više kupovala ponuda sa promocijskim paketom, dok kod pošte je više kupljenih sezonskih karata bez promo paketa. Kod ponuda koje su napravljene uživo u parku ostvaren je veliki broj prodaje i sa i bez promocijskog paketa, iako opet više bez promocijskog paketa. Zbog činjenice da način kontakta očito ima određenu ulogu u prodaji, potrebno je prilagoditi logističku regresiju:

```
Call:
glm(formula = as.factor(Pass) ~ Promo + Nacin, family = binomial,
    data = pass.df)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.07860    0.13167 -15.787  < 2e-16 ***
PromoSa_paketom   -0.56022    0.09031  -6.203 5.54e-10 ***
NacinPosta         2.17617    0.14651  14.854  < 2e-16 ***
NacinUzivo_u_parku 3.72176    0.15964  23.313  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4375.0  on 3155  degrees of freedom
Residual deviance: 3490.2  on 3152  degrees of freedom
AIC: 3498.2

Number of Fisher Scoring iterations: 4
```

Slika 10: Rezultati glm() funkcije

Temeljem rezultata glm() funkcije možemo primijetiti kako model procjenjuje dosta negativan utjecaj promocijskog paketa i dosta pozitivan utjecaj kanala pristupa putem pošte i uživo u parku. Na sljedećoj slici su prikazani omjeri izgleda i pouzdani intervali:

```
> exp(coef(pass.m2))
       (Intercept)    PromoSa_paketom         NacinPosta NacinUzivo_u_parku
         0.1251054          0.5710846          8.8125066         41.3371206
> exp(confint(pass.m2))
waiting for profiling to be done...
                       2.5 %     97.5 %
(Intercept)         0.09577568  0.1606189
PromoSa_paketom     0.47793969  0.6810148
NacinPosta          6.65770550 11.8328173
NacinUzivo_u_parku 30.42959274 56.9295369
```

Slika 11: Pouzdani intervali

U ovom modelu, uz prisutnost promocijskog paketa imamo 32-53% manju šansu za prodajom, dok s druge strane ponude kupcima uživo u parku nam povećavaju šansu za prodajom 30-56 puta. Ponovno se postavlja pitanje jesmo li uzeli u obzir sve moguće varijable te zbog toga nastavljamo dalje sa istraživanjem.

8.3 Model III.

Promatranjem podataka naišli smo na moguće efekte interakcije između promocijskog paketa i načina kojim se pristupilo kupcu. Ovo opažanje je najviše istaknuto na slici *doubledecker* plota na kojem se vidi drastična razlika koju efekt promocijskog paketa ima kada se ponudi putem e-maila. Zbog toga radimo novi model:

```
Call:
glm(formula = as.factor(Pass) ~ Promo + Nacin + Promo:Nacin,
    family = binomial, data = pass.df)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -2.8883     0.1977 -14.608  < 2e-16 ***
PromoSa_paketom                   2.1071     0.2783   7.571 3.71e-14 ***
NacinPosta                        3.1440     0.2133  14.743  < 2e-16 ***
NacinUzivo_u_parku                4.6455     0.2510  18.504  < 2e-16 ***
PromoSa_paketom:NacinPosta       -2.9808     0.3003  -9.925  < 2e-16 ***
PromoSa_paketom:NacinUzivo_u_parku -2.8115   0.3278  -8.577  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4375.0  on 3155  degrees of freedom
Residual deviance: 3393.5  on 3150  degrees of freedom
AIC: 3405.5

Number of Fisher Scoring iterations: 5
```

Slika 12: Rezultati glm() funkcije

Interakcija promocijskog paketa i načina ponude je statistički značajna i ima negativan efekt kod pošte i parka. Intervali su prikazani na sljedećoj slici:

```
                                    2.5 %       97.5 %
(Intercept)                    0.03688720   0.08032263
PromoSa_paketom                4.78970184  14.31465957
NacinPosta                    15.54800270  35.97860059
NacinUzivo_u_parku            64.74364028 173.57861021
PromoSa_paketom:NacinPosta     0.02795867   0.09102369
PromoSa_paketom:NacinUzivo_u_parku 0.03135437 0.11360965
```

Slika 13: Pouzdani intervali

Kroz omjere izgleda možemo vidjeti da je promocijski paket kroz poštu i park uspješan samo 2-11% u odnosu na email. Dakle, promocijski paket kroz e-mail ima najbolje rezultate te ga je i dalje dobro nuditi na taj način. Možemo zaključiti da uspjeh promocijskog

paketa ovisi o načinu ponude te da uspjeh prodaje preko emaila ne implicira uspjeh preko drugih kanala.

9 Primjer II.

Bank Marketing (https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing)
Sažetak: Podaci su povezani s izravnim marketinškim kampanjama (telefonskim pozivima) portugalske banke. Cilj klasifikacije je predvidjeti hoće li klijent sklopiti terminski depozit (varijabla y).
Informacije o skupu podataka: Podaci su povezani s izravnim marketinškim kampanjama portugalske banke. Marketinške kampanje su se temeljile na telefonskim pozivima. Često je bilo potrebno više kontakata s istim klijentom kako bi se saznalo hoće li se na proizvod pretplatiti (terminski depozit), opcije su binarne "Da" (yes) ili "Ne" (no).

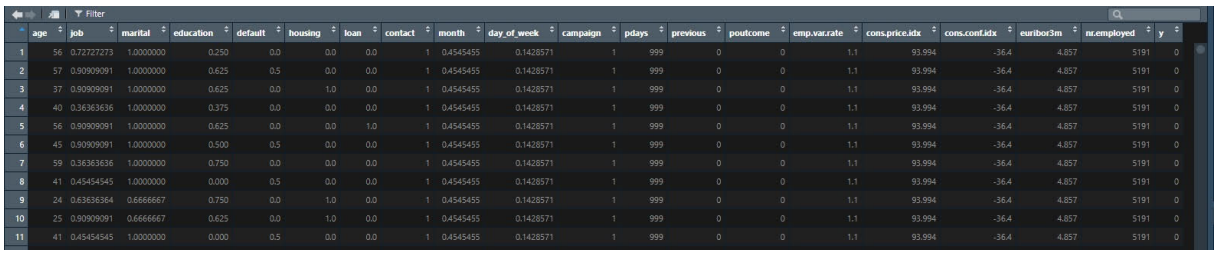
9.1 Informacije o atributima

1. Podaci o klijentima banke:
    - Dob: broj
    - Tip posla: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'.
    - Bračni status: 'divorced', 'married', 'single', 'unknown'.
    - Stupanj školovanja: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'.
    - Postojanje neizmirenih kredita: 'no', 'yes', 'unknown'.
    - Postojanje stambenog kredita: 'no', 'yes', 'unknown'.
    - Postojanje pozajmice (minus na računu i sl.): 'no', 'yes', 'unknown'.
2. Detalji o kontaktu tijekom trenutne kampanje:
    - Kontakt uspostavljen putem: 'mobitelom,'fiksnim telefonom'

- Mjesec u godini: 'jan', 'feb', 'mar', ..., 'nov', 'dec'.
- Dan u tjednu: 'mon','tue','wed','thu','fri'.

3. Ostali atributi:
   - Campaign: broj kontakata s tim klijentom u ovoj kampanji.
   - Pdays: broj dana proteklih od posljednjeg kontakta s klijentom iz prošle kampanje.
   - Previous: broj kontakata s klijentom u prethodnim kampanjama.
   - Poutcome: ishod u prošloj kampanji za tog klijenta: 'failure','nonexistent','success'.
4. Društveni i ekonomski atributi
   - Emp.var.rate: stopa promjene zaposlenosti - tromjesečni indikator.
   - Cons.price.idx: potrošački indeks cijena - mjesečni indikator.
   - Cons.conf.idx: indeks potrošačkog povjerenja - mjesečni indikator.
   - Euribor3m: euribor kamatna stopa na 3 mjeseca - dnevni indikator.
   - Nr.employed: broj zaposlenika - tromjesečni indikator.
5. Ciljna varijabla (željeni cilj):
   - y: je li klijent pretplaćen na terminski depozit? (binarno: 'yes', 'no').

9.2 Pregled dijela pripremljenih podataka



	age	job	marital	education	default	housing	loan	contact	month	day_of_week	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
1	56	0.72727273	1.0000000	0.250	0.0	0.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
2	57	0.90909091	1.0000000	0.625	0.5	0.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
3	37	0.90909091	1.0000000	0.625	0.0	1.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
4	40	0.36363636	1.0000000	0.375	0.0	0.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
5	56	0.90909091	1.0000000	0.625	0.0	0.0	1.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
6	45	0.90909091	1.0000000	0.500	0.5	0.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
7	59	0.36363636	1.0000000	0.750	0.0	0.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
8	41	0.45454545	1.0000000	0.000	0.5	0.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
9	24	0.63636364	0.6666667	0.750	0.0	1.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
10	25	0.90909091	0.6666667	0.625	0.0	1.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0
11	41	0.45454545	1.0000000	0.000	0.5	0.0	0.0	1	0.4545455	0.1428571	1	999	0	0	1.1	93.994	-36.4	4.857	5191	0

Slika 14: Pregled dijela pripremljenih podataka

Data set se sastojao od 41188 redova (klijenata) i 20 stupca (19 parametara + ishod). U našem primjeru nasumično smo izabrali 1000 klijenata za skup "Test" a preostalih 40-ak tisuća za "Train".

Dio koda u R-u:

$$broj\_redaka <- nrow(data)$$

$$test\_indices <- sample(1 : broj\_redaka, 1000)$$

$$test <- data[test\_indices,]$$

$$train <- data[-test\_indices,]$$

9.3 Model

Na setu podataka koristimo funkciju glm() iz R-a. Dobivamo sljedeće vrijednosti:

```
Call:
glm(formula = y ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9073  -0.3918  -0.3234  -0.2761   3.1443

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.256e+01  1.559e+01  -5.296 1.18e-07 ***
age             3.735e-03  1.545e-03   2.417 0.015635 *
job             5.114e-02  7.990e-02   0.640 0.522151
marital        -5.066e-02  5.614e-02  -0.902 0.366879
education       2.148e-01  7.841e-02   2.740 0.006145 **
default        -6.433e-01  1.141e-01  -5.638 1.72e-08 ***
housing        -1.086e-02  3.617e-02  -0.300 0.764094
loan           -3.773e-02  4.963e-02  -0.760 0.447151
contact        -1.080e+00  6.077e-02 -17.767  < 2e-16 ***
month          -7.035e-01  1.153e-01  -6.100 1.06e-09 ***
day_of_week     3.400e-01  8.751e-02   3.885 0.000102 ***
campaign       -4.441e-02  9.461e-03  -4.694 2.68e-06 ***
pdays          -1.871e-03  8.104e-05 -23.093  < 2e-16 ***
previous       -1.104e-01  5.380e-02  -2.052 0.040182 *
poutcome       -4.437e-01  8.389e-02  -5.289 1.23e-07 ***
emp.var.rate   -8.391e-01  6.305e-02 -13.309  < 2e-16 ***
cons.price.idx  1.137e+00  9.855e-02  11.537  < 2e-16 ***
cons.conf.idx   4.598e-02  5.619e-03   8.183 2.77e-16 ***
euribor3m       3.045e-01  9.073e-02   3.356 0.000790 ***
nr.employed    -4.452e-03  1.486e-03  -2.996 0.002736 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28252  on 40187  degrees of freedom
Residual deviance: 22557  on 40168  degrees of freedom
AIC: 22597

Number of Fisher Scoring iterations: 6
```

Slika 15: Rezultati glm() funkcije

Oznake $***$ otkrivaju koji su parametri najznačajniji prediktori uspješnosti prodaje bankarske usluge, a iz prikaza pouzdanih intervala vidimo koliki je utjecaj pomicanja unutar pojedine klase na vjerojatnost da će klijent odgovoriti pozitivno na ponudu banke.

```
                    2.5 %       97.5 %
(Intercept)    8.381142e-50 2.918471e-23
age            1.000703e+00 1.006783e+00
job            8.996712e-01 1.230567e+00
marital        8.520560e-01 1.061839e+00
education      1.063519e+00 1.446234e+00
default        4.193641e-01 6.559573e-01
housing        9.215283e-01 1.061930e+00
loan           8.731512e-01 1.060705e+00
contact        3.013774e-01 3.824458e-01
month          3.945759e-01 6.201527e-01
day_of_week    1.183541e+00 1.667902e+00
campaign       9.386129e-01 9.740664e-01
pdays          9.979710e-01 9.982882e-01
previous       8.059837e-01 9.953485e-01
poutcome       5.438838e-01 7.557284e-01
emp.var.rate   3.818459e-01 4.889088e-01
cons.price.idx 2.568193e+00 3.779311e+00
cons.conf.idx  1.035591e+00 1.058655e+00
euribor3m      1.135139e+00 1.620022e+00
nr.employed    9.926534e-01 9.984533e-01
```

Slika 16: Pouzdani intervali

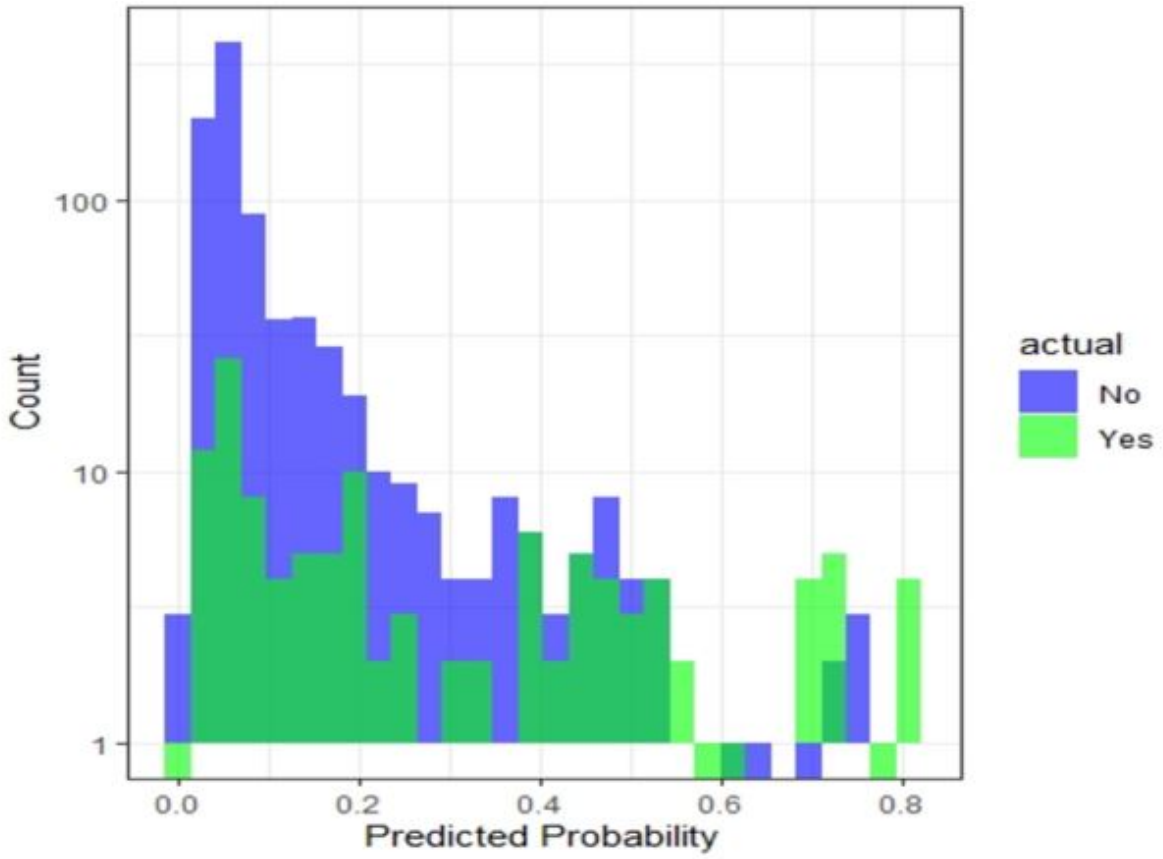
Model je bio precizan u oko 89% slučajeva, a iz prikaza pomoću konfuzijske matrice možemo vidjeti kako je pogađao:

```
> confusion_matrix <- table(predv, test$y)
>
> print(confusion_matrix)

predv   0   1
    0 862  99
    1  15  24
>
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
> print(paste("Točnost (Accuracy):", accuracy))
[1] "Točnost (Accuracy): 0.886"
>
```

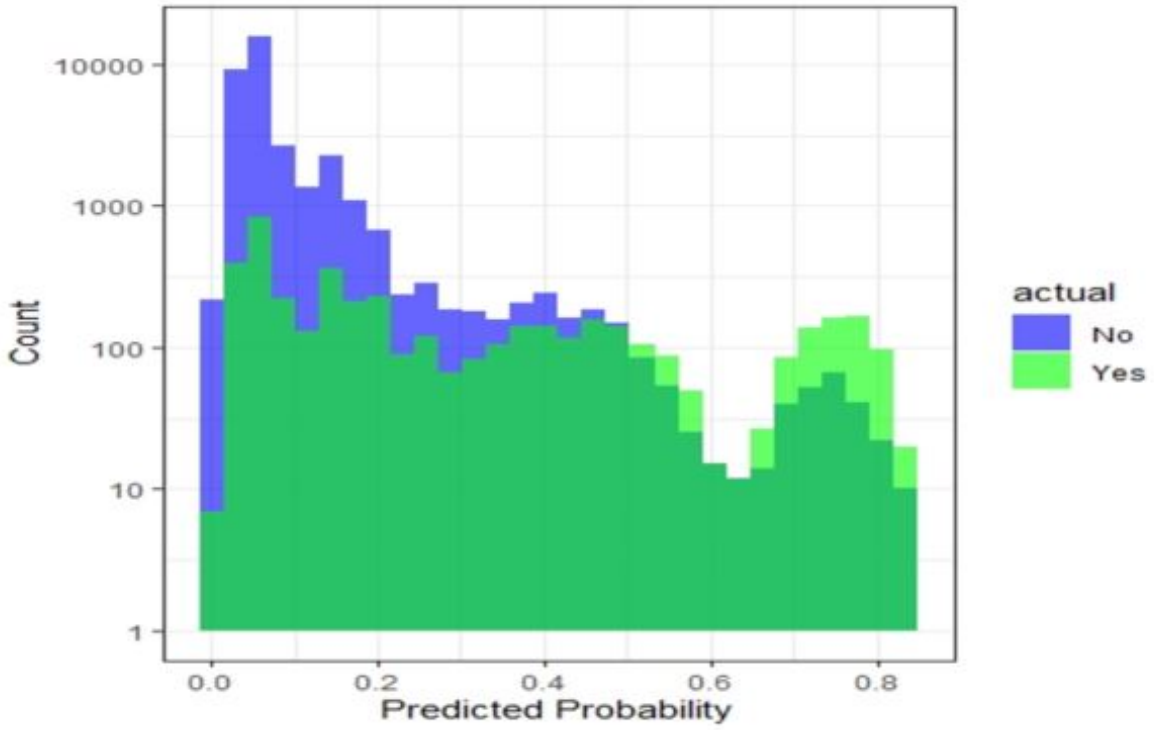
Slika 17: Konfuzijska matrica

Dakle, u ovom nasumično odabranom uzorku predviđali smo da klijent neće koristiti uslugu s točnošću $\frac{862}{862+99} = 0.897$, a da će koristiti uslugu s točnošću $\frac{24}{24+15} = 0.615$. Možda još vrijedi pogledati i grafički prikaz testnog uzorka, koji zelenim stupcima prikazuje logaritam vrijednosti broja klijenata koji su pristali na ponudu banke, te plavom one koji nisu pristali, a za koje je model logističke regresije procijenio vjerojatnost pozitivnog odgovora unutar vrijednosti intervala na apscisi.



Slika 18: Grafički prikaz testnog uzorka

Naš model postavljen je tako da vjerojatnosti iznad 50% označava kao "predviđanje uspjeha - yes", a one ispod kao "predviđanje neuspjeha - no". Na 50% smo se odlučili nakon promatranja istog tipa prikaza za podatke "Train":



Slika 19: Grafički prikaz

9.4 Izvedeni zaključak

Precision (preciznost) je mjera koja se koristi u evaluaciji performansi klasifikacijskih modela, posebno u kontekstu problema binarne klasifikacije. Precision se definira kao omjer broja pravilno pozitivno klasificiranih instanci (true positives) i ukupnog broja pozitivno klasificiranih instanci, što se može izraziti formulom: $Precision = \frac{truepositives}{truepositives + falsepositives}$ (62% u našem primjeru). Precision mjeri koliko je točan model u klasifikaciji pozitivnih instanci. Visoka preciznost ukazuje na to da model ima malo lažno pozitivnih predviđanja, odnosno da je vrlo pouzdan u prepoznavanju pozitivnih instanci. S druge strane, niska preciznost ukazuje na veći broj lažno pozitivnih predviđanja. Precision je važan kada je potrebno minimizirati lažno pozitivne predikcije, tj. kada je bitno da pozitivno klasificirane instance budu što preciznije identificirane.

Recall (odziv), također poznat kao osjetljivost, se definira kao omjer broja pravilno pozitivno klasificiranih instanci (true positives) i ukupnog broja stvarno pozitivnih instanci, što se može izraziti formulom: $Recall = \frac{truepositives}{truepositives + falsenegatives}$ (20% u našem primjeru).

Recall mjeri koliko je dobar model u klasifikaciji pozitivnih instanci i koliko uspješno otkriva sve stvarno pozitivne instance. Visoki recall ukazuje na to da model ima malo lažno negativnih predviđanja, odnosno da ima sposobnost identificirati većinu pozitivnih instanci. S druge strane, niski recall ukazuje na veći broj lažno negativnih predviđanja. Recall je važan kada je prioritet prepoznavanje što većeg broja stvarno pozitivnih instanci, bez obzira na broj lažno pozitivnih predikcija.

Važno je napomenuti da precision ne uzima u obzir lažno negativne predikcije, odnosno ne mjeri točnost u identifikaciji negativnih instanci, dok s druge strane recall ne uzima u obzir lažno pozitivne predikcije, odnosno ne mjeri točnost u identifikaciji negativnih instanci.

Za cjelovitu procjenu performansi modela kombinira se preciznost i odziv.tj. F1-mjera, koja ih kombinira kako bi pružila cjelovitu procjenu performansi klasifikacijskog modela, posebno u kontekstu problema binarne klasifikacije. F1-mjera je harmonička sredina preciznosti i odziva, a izračunava se formulom: $F1 = 2 * \frac{precision * recall}{precision + recall}$ (30% u našem primjeru). Ona pruža balansiranu procjenu performansi modela tako što uzima u obzir i lažno pozitivne i lažno negativne predikcije. Visoka F1-mjera ukazuje na to da model ima dobru ravnotežu između preciznosti i odziva, odnosno dobro klasificira pozitivne instance i minimizira lažno pozitivne i lažno negativne predikcije. Ta mjera je korisna kada je važno postići balans između preciznosti i odziva, a ne preferirati samo jedan od tih aspekata. Na primjer, ako je važno pravilno klasificirati pozitivne instance, ali istovremeno minimizirati lažno pozitivne predikcije i lažno negativne propuste, F1-mjera će pružiti cjelovitu procjenu performansi modela. U našem primjeru fokus je bio više usmjeren prema prepoznavanju mogućih kupaca, pa smo tako i modelirali logističku regresiju.

Literatura

[1] https://www.ibm.com/topics/logistic-regression

[2] https://hr.theastrologypage.com/logistic-regression

[3] https://www.statisticshowto.com/hosmer-lemeshow-test/

[4] https://online.stat.psu.edu/stat462/node/207/

[5] https://bookdown.org/jarneric/predavanja_smea/2-2-logit-transformacija.html

[6] https://www.ibm.com/topics/logistic-regression?mhsrc=ibmsearch_amhq=logistic%20regression

[7] https://peopleanalytics-regression-book.org/bin-log-reg.html