Modeliranje odljeva igrača u online klađenjima primjenom neuronskih mreža i logističke regresije

Crnobrnja, Martina

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku

Martina Crnobrnja

Modeliranje odljeva igrača u online klađenjima primjenom neuronskih mreža i logističke regresije

Diplomski rad

Osijek, 2020.

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku

Martina Crnobrnja

Modeliranje odljeva igrača u online klađenjima primjenom neuronskih mreža i logističke regresije

Diplomski rad

Mentor: prof.dr.sc. Nataša Šarlija
Komentor: doc.dr.sc. Slobodan Jelić

Osijek, 2020.

Sadržaj

1 Neuronske mreže **2**
- 1.1 Struktura neuronske mreže . . . 2
- 1.2 Učenje neuronske mreže . . . 4
- 1.3 Određivanje optimalnih parametara . . . 5
- 1.4 Propagacija unatrag . . . 5
- 1.5 Arhitektura neuronskih mreža primijenjana u istraživanju i način odabira modela 6
  - 1.5.1 Matrica konfuzije . . . 7
  - 1.5.2 ROC krivulja . . . 8
  - 1.5.3 Površina ispod ROC krivulje . . . 8

2 Logistička regresija **10**
- 2.1 Linearna regresija . . . 10
- 2.2 Višestruka linearna regresija . . . 12
- 2.3 Eksponencijalna familija distribucija i generalizirani linearni modeli . . . 13
- 2.4 Model logističke regresije . . . 16
- 2.5 Model logističke regresije primijenjen u istraživanju i način odabira modela . . . 18
  - 2.5.1 Anova . . . 18
  - 2.5.2 AIC . . . 18

3 Analiza zavisnosti **19**
- 3.1 Pearsonov χ^2 test nezavisnosti za diskretne slučajne varijable . . . 19
- 3.2 t-test za testiranje jednakosti očekivanja . . . 19

4 Empirijsko istraživanje: Modeliranje odljeva igrača u online klađenjima **20**
- 4.1 Prethodna istraživanja . . . 20
- 4.2 Varijable i podaci . . . 20
- 4.3 Model neuronskih mreža za procjenu odljeva igrača u online klađenjima . . . 31
- 4.4 Model logističke regresije za procjenu odljeva igrača u online klađenjima . . . 34
- 4.5 Usporedba modela neuronske mreže i logističke regresije za procjenu vjerojatnosti odljeva igrača . . . 37

5 Zaključak **38**

Popis slika **39**

Popis tablica **39**

Literatura **41**

Uvod

Cjelokupna djelatnost kockanja i klađenja može se promatrati prema vrstama igara na sreću, a to su lutrijske igre, igre u casinima, igre klađenja i igre na automatima. Prema Europskom udruženju za klađenje i igre na sreću[1], online klađenje i igre na sreću čine 20,7% ukupnog tržišta Europske unije. Klađenje na sport je najpopularniji oblik online klađenja u EU (40,3%), nakon čega slijede igre na sreću u casinima (32,1%) i lutrijske igre (13,3%). Na rast tržišta u Hrvatskoj osim promjene poreznog modela u 2014. godini velik utjecaj ima i razvoj digitalizacije, posebice u dijelu online igara na sreću. Tržište, osim profitabilnosti, karakterizira i izrazita koncentracija, koja je djelomice rezultat pripajanja i akvizicija te ulaska inozemnog kapitala (vidi [13]).

Ključno je zadržati postojeće klijente jer je stjecanje novih klijenata skuplje i dugotrajnije od zadržavanja starih. Štoviše, dugogodišnji klijenti troše više od novih klijenata. Istraživanje Earla Sassera i tvrtke Bain & Company[2] pokazalo je da povećanje stope zadržavanja klijenata za 5% može dovesti do povećanja dobiti od 25 do 95% (vidi [11]). S druge strane, zadržavanje klijenata sa sobom nosi troškove i teško je imati takve troškove za svakog klijenta, tj. nije svaki klijent vrijedan zadržavanja.

Cilj ovog rada je modelirati odljev klijenata na tržištu igara na sreću pomoću logističke regresije i neuronskih mreža te usporediti dobivene rezultate. Predviđanje odljeva klijenata je binarna klasifikacija čiji je rezultat vjerojatnost odljeva. U ovom radu ćemo napraviti modele za predikciju vjerojatnosti odljeva pomoću kojih će se identificirati oni klijenti koji će značajno smanjiti ili u potpunosti prestati s kockanjem kod priređivača igara na sreću. Priređivač tada može takvim klijentima ponuditi različite proizvode i nagrade s ciljem sprječavanja njihovog odlaska. Klijente koji će nastaviti s klađenjem nazivat ćemo aktivnima, a klijente koji su se prestali kladiti neaktivnima. Predviđanjem neaktivnih igrača želimo smanjiti odljev klijenata.

U prvom poglavlju obrađen je model neuronske mreže. Opisana je njena struktura, učenje, određivanje optimalnih parametara i propagacija unatrag, nakon čega je ukratko opisano na koji način je nuronska mreža primjenjena u istraživanju.

U drugom poglavlju obrađen je model logističke regresije. Polazeći od linearne regresije, eksponencijalnih familija distribucija i generaliziranih linearnih modela dolazimo do pojmova logističke regresije i opisujemo na koji je način logistička regresija primjenjena u modelu.

U trećem poglavlju opisan je način testiranja nezavisnosti između dvije varijable.

U četvrtom poglavlju modelira se odljev klijenata. Prvo su navedena prethodna istraživanja na temu odljeva klijenata. Nakon toga objašnjeni su podaci i dan je profil neaktivnog igrača. Opisan je način izrade modela neuronskih mreža i logističke regresije te je testirana kvaliteta modela. Uspoređeni su rezultati dobiveni modelom neuronske mreže i logističke regresije. Na kraju su navedeni zaključci dobiveni u istraživanju.

[1] European gaming and betting association(EGBA)

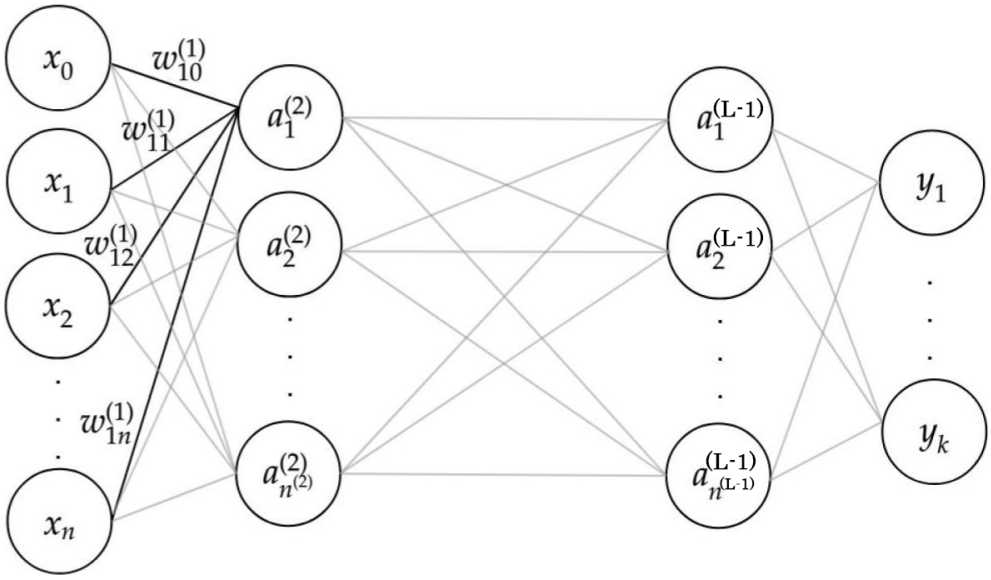
[2] Bain & Company je poduzeće za svjetovanje menadžmenta.

1 Neuronske mreže

Neuronske mreže su trenutno jedna od najpopularnijih i najkorištenijih metoda strojnog učenja. Temu neuronskih mreža započeli su Warren McCulloch i Walter Pitts 1943. u svom radu *A Logical Calculus of Ideas Immanent in Nervous Activity*. Neuronske mreže koriste se za rješavanje problema klasifikacije, regresije i predikcije u različitim područjima poput identifikacije, prepoznavanja uzoraka, medicinske dijagnostike, financijama i mnogim drugim zbog svoje sposobnosti da modeliraju nelinearne procese.

1.1 Struktura neuronske mreže

Umjetna neuronska mreža je metoda strojnog učenja inspirirana načinom učenja u prirodnoj neuronskoj mreži koja se nalazi u mozgu. Sastavljena je od slojeva koji se mogu podijeliti u tri osnovne skupine: ulazni sloj, skriveni sloj i izlazni sloj. Ulazni sloj neuronske mreže čine ulazni podaci, izlazni sloj posljednji je sloj mreže koji sadrži izlazne vrijednosti, a svi slojevi između ulaznog i izlaznog sloja nazivaju se skrivenim slojevima. Svaki sloj sastavljen je od aktivacijskih jedinica. Aktivacijska jedinica u skrivenom sloju linearna je kombinacija jedinica iz prethodnih slojeva. Svakoj aktivacijskoj jedinici pridružena je aktivacijska funkcija.



Slika 1: Struktura neuronske mreže koja ulazni podatak veličine n transformira kroz L slojeva i pridružuje predikcije za pripadanje svakoj od k kategorija

Konstruirajmo $n^{(2)}$ linearnih kombinacija ulaznih podataka $x_1, x_2, \ldots, x_n$ u sljedećem obliku

$$a_j^{(2)} = \sum_{i=1}^{n} w_{ji}^{(1)} x_i + w_{j0}^{(1)},$$

$j = 1, \ldots, n^{(2)}$. Parametri $w_{ji}^{(1)}$ su težine koje pripadaju prvom sloju, a parametar $w_{j0}^{(1)}$ je pomak (*engl.* bias). Veličine a_j zovu se aktivacijske jedinice. Svaka izračunata aktivacijska jedinica

se transformira nekom diferencijabilnom, nelinearnom funkcijom h koja se naziva aktivacijska funkcija na sljedeći način

$$z_j^{(2)} = h(a_j^{(2)}).$$

Pomoću tako dobivenih vrijednosti $z_j^{(2)}$ računaju se aktivacijske jedinice u sljedećem sloju. Općenito je k-ta aktivacijska jedinica u l-tom sloju linearna kombinacija vrijednosti aktivacijskih funkcija iz sloja $l-1$,

$$a_k^{(l)} = \sum_{j=1}^{n^{(l-1)}} w_{kj}^{(l-1)} z_j^{(l-1)} + w_{k0}^{(l-1)},$$

$k = 1, \ldots, n^{(l)}$, gdje $n^{(l)}$ predstavlja broj aktivacijskih jedinica u l-tom sloju. Uobičajeno je uvesti dodatnu varijablu $x_0 := 1$ tako da se pomak apsorbira u vektor težina $\mathbf{w}$ pa aktivacijsku jedinicu u sloju l možemo zapisati kao

$$a_k^{(l)} = \sum_{i=0}^{n^{(l-1)}} w_{ki}^{(l-1)} z_i^{(l-1)}.$$

U izlaznom sloju aktivacijske jedinice se preslikavaju u izlazne vrijednosti y_k pomoću prikladne aktivacijske funkcije za izlazni sloj, tj.

$$y_k = \tilde{g}(a_k^{(L)}) = \tilde{g}(\sum_{i=0}^{n^{(L-1)}} w_{ki}^{(L-1)} z_i^{(L-1)}),$$

$k = 1, \ldots, n^{(L)}$, gdje $n^{(L)}$ predstavlja broj izlaznih vrijednosti. Koristimo notaciju $\tilde{g}$ za aktivacijsku funkciju u izlaznom sloju kako bismo naglasili da aktivacijska funkcija u izlaznom sloju ne mora biti ista kao aktivacijska funkcija u skrivenim slojevima. Izbor aktivacijske funkcije u izlaznom sloju ovisi o podacima i problemu koji modeliramo.

Aktivacijska funkcija u skrivenim slojevima najčešće se bira između jedne od sljedećih funkcija:

- sigmoidna (logistička) funkcija

$$\sigma : \mathbb{R} \to [0,1], \qquad \sigma(a) = \frac{1}{1+\mathrm{e}^{-a}},$$

- tangens hiperbolni

$$\tanh : \mathbb{R} \to [-1,1], \qquad \tanh(a) = \frac{\sinh(a)}{\cosh(a)} = \frac{\mathrm{e}^{2a}-1}{\mathrm{e}^{2a}+1},$$

- rectified linear unit ($ReLU$)

$$ReLU : \mathbb{R} \to [0,\infty], \qquad ReLU(a) = max\{0,a\}.$$

1.2 Učenje neuronske mreže

Pretpostavimo da je zadan slučaj binarne klasifikacije s jednom izlaznom varijablom t. Za ulazni podatak $\mathbf{x}$ za koji je izlazna vrijednost $t = 1$ kažemo da pripada klasi C_1, inače mu je izlazna vrijednost $t = 0$ i pripada klasi C_0. Pretpostavimo da je izlazna vrijednost dobivena sigmoidnom aktivacijskom funkcijom u izlaznom sloju, tj. izlazna vrijednost s obzirom na aktivacijsku jedinicu a je

$$y(\mathbf{x}, \mathbf{w}) = \sigma(a) = \frac{1}{1 + \mathrm{e}^{-a}}$$

i vrijedi $0 \leq y(\mathbf{x}, \mathbf{w}) \leq 1$. Vrijednost $y(\mathbf{x}, \mathbf{w})$ može se interpretirati kao uvjetna vjerojatnost $p(C_1|\mathbf{x})$, dok je $p(C_0|\mathbf{x})$ dana s $1 - y(\mathbf{x}, \mathbf{w})$. Dakle, uvjetne distribucije izlaznih varijabli modelirane su Bernoullijevom distribucijom s mogućim realizacijama iz skupa $\{0, 1\}$ i vjerojatnostima

$$p(t|\mathbf{x}, \mathbf{w}) = y(\mathbf{x}, \mathbf{w})^t (1 - y(\mathbf{x}, \mathbf{w}))^{1-t}.$$

Označimo $y = y(\mathbf{x}, \mathbf{w})$. Djelovanjem negativnog logaritma na prethodni izraz slijedi

$$-\ln(p(t|\mathbf{x}, \mathbf{w})) = -t \ln(y) - (1 - t) \ln(1 - y).$$

Ako su podaci u skupu za treniranje nezavisni uzorci, tada se funkcija greške dobivena pomoću gore navedenih log likelihood funkcija naziva funkcija unakrsne entropije (*engl.* cross-entropy function) i ima sljedeći oblik

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\},$$

gdje y_n označava $y(\mathbf{x}_n, \mathbf{w})$, a N broj ulaznih podataka.

Ako je potrebno ulazne podatke klasificirati u $K > 2$ klasa, koristi se mreža s K jedinica u izlaznom sloju. Svakoj od K aktivacijskih jedinica u izlaznom sloju pridružena je vrijednost $t_k \in \{0, 1\}$, tj. $t_i = 1$ označava pripadnost ulaznog podatka i-toj klasi i tada je $t_j = 0, \forall j \neq i$. Izlazne vrijednosti mreže mogu se interpretirati kao $y_k(\mathbf{x}, \mathbf{w}) = p(t_k = 1|\mathbf{x})$ iz čega slijedi sljedeća funkcija greške

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{kn} \ln y_k(\mathbf{x}_n, \mathbf{w}).$$

Na aktivacijske jedinice u izlaznom sloju djeluje se sa softmax aktivacijskom funkcijom kako bi se dobile izlazne vrijednosti $y_k(\mathbf{x}, \mathbf{w})$,

$$y_k(\mathbf{x}, \mathbf{w}) = softmax(a)_k = \frac{\mathrm{e}^{a_k}}{\sum_{j=1}^{K} \mathrm{e}^{a_j}}.$$

Može se pokazati da vrijedi $0 \leq y_k \leq 1$ i $\sum_{k=1}^{K} y_k = 1$.

1.3 Određivanje optimalnih parametara

Određivanje vektora težina $\mathbf{w}$ problem je minimizacije funkcije greške $E(\mathbf{w})$. Funkcije greške iz prethodnog poglavlja su diferencijabilne funkcije na kojima se mogu primjeniti iterativne metode iz numeričke matematike.

Iterativni proces za traženje točke lokalnog minimuma dovoljno puta neprekidno diferencijabilne funkcije $f : \mathbb{R}^n \to \mathbb{R}$ oblika je

$$x_{k+1} = x_k + \alpha_k p_k, \qquad k = 0, 1, \ldots,$$

gdje je x_0 početna aproksimacija, p_k vektor smjera kretanja od točke x_k u točku x_{k+1}, a $\alpha_k > 0$ duljina koraka u smjeru vektora p_k. Kretanje od točke x_k u točku x_{k+1} treba ostvariti tako da se postigne smanjenje vrijednosti funkcije f, tj. tako da vrijedi $f(x_{k+1}) < f(x_k)$. To se može postići odabirom vektora p_k tako da vrijedi

$$(\nabla f(x_k), p_k) < 0,$$

gdje $\nabla f(x_k)$ označava gradijent funkcije f u točki x_k (vidi [12]).

Obična gradijentna metoda za minimizaciju funkcije $f : \mathbb{R}^n \to \mathbb{R}$ za vektor smjera kretanja uzima $p_k = -\nabla f(x_k)$ koji zadovoljava uvjet $(\nabla f(x_k), p_k) < 0$. U praksi se pri treniranju neuronskih mreža na velikim bazama podataka korisnom pokazala metoda stohastičkog gradijentnog spusta (*engl.* stochastic gradient descent). Stohastički gradijentni spust korigira težine nakon izračuna iteracije za svaki ulazni podatak

$$w^{(k+1)} = w^{(k)} - \eta \nabla E_n(w^{(k)}).$$

Funkcija greške definirana pomoću metode maksimalne vjerodostojnosti za skup nezavisnih uzoraka može se računati kao zbroj vrijednosti po podacima, tj.

$$E(\mathbf{w}) = \sum_{n=1}^{N} E_n(\mathbf{w}).$$

1.4 Propagacija unatrag

Sada treba pronaći učinkovitu tehniku za procjenu gradijenta funkcije greške $E(\mathbf{w})$. Za zadani ulaz $\mathbf{x}_n$ računaju se vrijednosti aktivacijskih jedinica

$$a_k^{(l)} = \sum_{i=0}^{n^{(l-1)}} w_{ki}^{(l-1)} z_i^{(l-1)} \tag{1.1}$$

na koje onda djeluje aktivacijska funkcija h tako da daje vrijednosti $z_j^{(l)} = h(a_j^{(l)})$.

Pretpostavimo da smo za svaki ulazni podatak izračunali aktivacijske jedinice u svim skrivenim i izlaznom sloju. Taj postupak naziva se propagacija unaprijed (*engl.* forward propagation) jer tijek informacija ide prema naprijed kroz mrežu. Sada treba pronaći parcijalne derivacije funkcija $E_n(\mathbf{w})$ s obzirom na težinu $w_{ji}^{(l)}$. Funkcija E_n ovisi o težini $w_{ji}^{(l)}$ samo preko aktivacijske jedinice $a_j^{(l+1)}$, pa možemo primijeniti lančano pravilo

$$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \frac{\partial E_n}{\partial a_j^{(l+1)}} \frac{\partial a_j^{(l+1)}}{\partial w_{ji}^{(l)}}. \tag{1.2}$$

Označimo s

$$\delta_j^{(l)} := \frac{\partial E_n}{\partial a_j^{(l+1)}}.$$

Iz (1.1) slijedi

$$\frac{\partial a_j^{(l+1)}}{\partial w_{ji}^{(l)}} = z_i^{(l)}.$$

Sada (1.2) možemo zapisati u sljedećem obliku

$$\frac{\partial E_n}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} z_i^{(l)}.$$

Sada je potrebno izračunati $\delta_j^{(l)}$ za $l = 1, \ldots, L-1$, pri čemu L označava ukupan broj slojeva u neuronskoj mreži. Za vrijednosti u izlaznom sloju vrijedi

$$\delta_k^{(L)} = y_k - t_k,$$

a t_k je stvarna vrijednost modelirane varijable. Općenito za procjenu $\delta_j^{(l)}$ iz skrivenog sloja primjenom lančanog pravila vrijedi

$$\delta_j^{(l)} = \frac{\partial E_n}{\partial a_j^{(l+1)}} = \sum_k \frac{\partial E_n}{\partial a_k^{(l+2)}} \frac{\partial a_k^{(l+2)}}{\partial a_j^{(l+1)}}, \tag{1.3}$$

gdje suma ide po svim aktivacijskim jedinicama $a_k^{(l+2)}$ koje u svojoj linearnoj kombinaciji sadrže transformaciju od $a_j^{(l+1)}$. Kako za općenito za aktivaciju u sloju l vrijedi

$$a_k^{(l)} = \sum_{j=0}^{n^{(l-1)}} w_{kj}^{(l-1)} z_j^{(l-1)} = \sum_{i=0}^{n^{(l-1)}} w_{ki}^{(l-1)} h(a_i^{(l-1)}),$$

uočavamo

$$\frac{\partial a_k^{(l+2)}}{\partial a_j^{(l+1)}} = w_{kj}^{(l+1)} h'(a_j^{(l+1)}).$$

Uvrštavanjem prethodne jednakosti u (1.3) slijedi formula

$$\delta_j^{(l)} = h'(a_j^{(l+1)}) \sum_k w_{kj}^{(l+1)} \delta_k^{(l+1)}. \tag{1.4}$$

Parcijalne derivacije računaju se od zadnjeg sloja prema natrag pa se formula (1.4) naziva propagacija unatrag (*engl.* backpropagation).

1.5 Arhitektura neuronskih mreža primijenjana u istraživanju i način odabira modela

Za kreiranje neuronskih mreža korištena je besplatna programska knjižnica Scikit-learn koja implementira širok spektar funkcija strojnog učenja u programskom jeziku Python. Algoritam korišten u ovom radu je Multi-layer Perceptron (MLP) i opisan je ranije u ovom poglavlju. Jedna od prednosti MLP-a je sposobnost učenja nelinearnog modela. Neki nedostatci MLP-a uključuju:

- MLP sa skrivenim slojevima ima nekonveksnu funkciju gubitka koja ima više lokalnih minimuma. Stoga drukčije početne težine mogu dovesti do različitih točnosti provjere valjanosti.
- MLP zahtjeva prilagođavanje broja parametara kao što su broj skrivenih slojeva, čvorova u skrivenim slojevima i broj iteracija.
- MLP je osjetljiv na neskalirane podatke.

Klasa MLPClassifier implementira algoritam višeslojnog perceptrona koji trenira pomoću propagacije unatrag. Nakon treniranja model može predviđati klasu za nove podatke iz skupa za testiranje. Trenutno MLPClassifier podržava jedino funkciju unakrsne entropije kao funkciju gubitka. Algoritam staje kada dosegne unapijed zadani maksimalni broj iteracija ili kada poboljšanje funkcije gubitka postane manje od nekog unaprijed zadanog broja $tol = 0,0001$.

Za validaciju modela neuronskih mreža koristit ćemo se matricom konfuzije i ROC krivuljama koje definiramo u nastavku.

1.5.1 Matrica konfuzije

Nakon definiranja modela potrebno je testirati koliko dobro model klasificira klijente u kategorije.

Označimo s a_{11} broj točno predviđenih neaktivnih klijenata (TP, *engl.* true positive), a_{12} broj netočno predviđenih neaktivnih klijenata (FN, *engl.* false negative), a_{21} broj netočno predviđenih aktivnih klijenata (FP, *engl.* false positive) i a_{22} broj točno predviđenih aktivnih klijenata (TN, *engl.* true negative). Tablica 1 prikazuje matricu konfuzije.

	Predviđeni aktivni	Predviđeni neaktivni
Stvarno aktivni	a_{11}	a_{12}
Stvarno neaktivni	a_{21}	a_{22}

Tablica 1: Matrica konfuzije

Iz matrice konfuzije definiramo senzitivnost(TPR) kao udio točno predviđenih aktivnih klijenata od ukupnog broja stvarno aktivnih klijenata, tj.

$$TPR = \frac{a_{11}}{a_{11} + a_{12}}.$$

Stopu lažno pozitivnih definiramo kao $FPR = 1 - TNR$, pri čemu je

$$TNR = \frac{a_{22}}{a_{21} + a_{22}}.$$

TNR (*engl.* true negative rate) još nazivamo specifičnost (*engl.* specificity). Vrijedi:

$$FNR = \frac{a_{21}}{a_{21} + a_{22}}$$

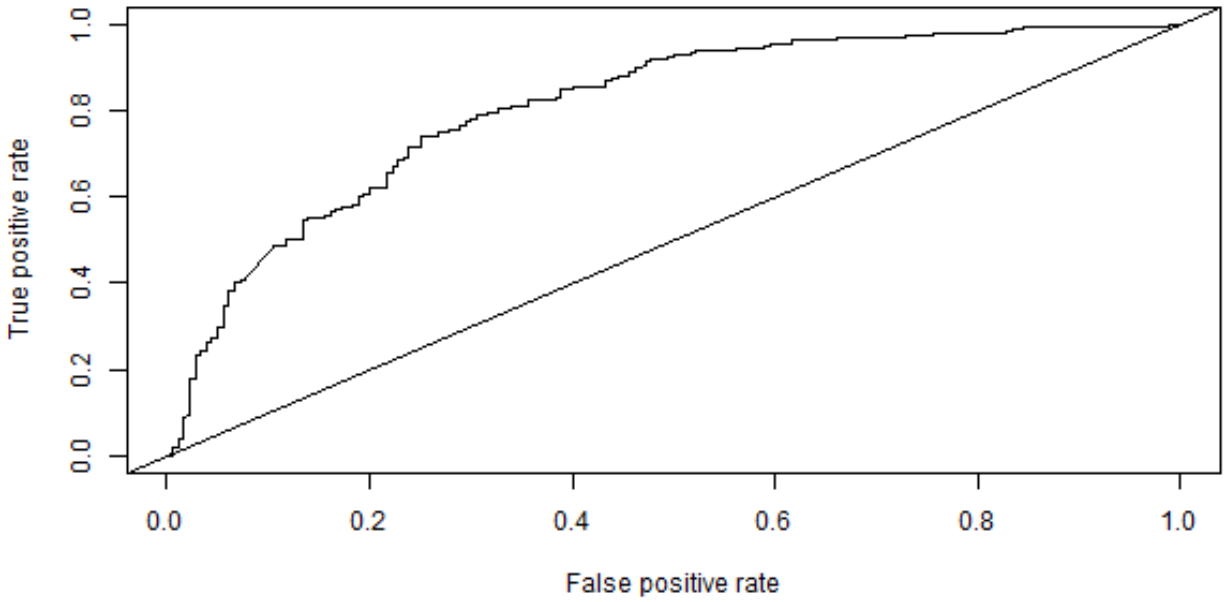
je udio stvarno neaktivnih klijenata koji su pogrešno predviđeni kao aktivni u ukupnom broju stvarno neaktivnih klijenata.

Dodatno još definiramo točnost(*engl.* accuracy) kao udio točno predviđenih u ukupnom broju klijenata

$$ACC = \frac{a_{11} + a_{22}}{a_{11} + a_{12} + a_{21} + a_{22}}.$$

1.5.2 ROC krivulja

Jedan od čestih načina ocjenjivanja klasifikacijske sposobnosti modela je ROC krivulja (*engl.* receiver operating characteristic curve). ROC krivulja je graf koji prikazuje kumulativne frekvencije loših(neaktivnih) klijenata na ordinati i kumulativne frekvencije dobrih klijenata na apcisi. Idealan model bi bio onaj koji sve loše klijente točno klasificira kao loše (vidi [17]). Na slici 2 prikazan je primjer grafa ROC krivulje.



Slika 2: Primjer ROC krivulje

1.5.3 Površina ispod ROC krivulje

AUC (*eng.* Area under Curve) je površina područja ispod ROC krivulje. Veća AUC vrijednost ukazuje na bolju klasifikacijsku moć modela. AUC vrijednost veću od 0,7 imaju dobro prilagođeni modeli, a AUC vrijednost veću od 0,8 imaju izvrsno prilagođeni modeli. Za modele s AUC vrijednosti manjom od 0,5 kažemo da su loše prilagođeni. Model kojemu je AUC jednak

0,5 ne predviđa bolje od modela slučajnog izbora (npr. bacanje pravilnog novčića). Usporedbom dva modela samo na osnovi AUC vrijednosti nije odmah jasno koji je od njih bolji jer ne znamo kako izgledaju njihove ROC krivulje. Osim geometrijske interpretacije, AUC vrijednost možemo interpretirati kao vjerojatnost da će loš klijent nasumično odabran iz uzorka imati lošiji rejting od nasumično izabranog dobrog klijenta (vidi [17]).

2 Logistička regresija

Pojam *regresije* uveo je Sir Francis Galton[3] kako bi opisao vezu između visine očeva i visine sinova. Uočio je kako će visina sina biti između očeve visine i prosječne visine. Ovaj efekt nazvao je *regresija prema prosjeku*. Regresijska analiza koristi se za donošenje zaključaka o slučajnoj varijabli Y ili o nizu slučajnih varijabli $Y_1, \ldots, Y_n$ koje ovise o nezavisnoj varijabli x ili vektoru varijabli $\mathbf{x} = (x_0, x_1, \ldots, x_p)$ koje nazivamo prediktorima.

2.1 Linearna regresija

Model čije je očekivanje linearna funkcija parametara naziva se model linearne regresije. Očekivanje zavisne varijable Y modeliramo pomoću nezavisnih varijabli $x_0, x_1, \ldots, x_p$ u obliku

$$\mathrm{E}(Y) = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \tag{2.1}$$

gdje su $\beta_0, \beta_1, \ldots, \beta_p$ nepoznati parametri. Dodatno pretpostavljamo da varijanca ne ovisi o $\mathbf{x}$, tj. $Var(Y) = \sigma^2$, te da su vrijednosti $x_0, x_1, \ldots, x_p$ izmjerene bez greške. Model se može zapisati i na sljedeći način

$$Y = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon, \tag{2.2}$$

gdje je ϵ slučajna varijabla takva da je $\mathrm{E}(\epsilon) = 0$ i $Var(\epsilon) = \sigma^2$. Jedan od zadataka linearne regresije je procjena nepoznatih parametara $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ radi određivanja predikcija vrijednosti ovisne varijable i interpretacije veze ovisne varijable s prediktorima. U jednadžbu je moguće uvesti konstantni član postavljanjem prvog elementa u $\mathbf{x}$ na 1. Za $\mathbf{x} = (1, x_1, \ldots, x_p)$ vrijedi $\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. Najprije ćemo promotriti jednostavni oblik linearne regresije za koji je $p = 1$ i $\mathbf{x} = (1, x_1)$.

Model oblika

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{2.3}$$

gdje je $\mathrm{E}(\epsilon) = 0$ i $Var(\epsilon) = \sigma^2$, naziva se modelom jednostavne linearne regresije. Procijenimo sada vrijednost parametara β_0 i β_1 metodom najmanjih kvadrata(*engl.* Least Square Method).

Pretpostavimo da su $x_0, x_1, \ldots, x_n$ realni brojevi koji daju opažene vrijednost niza n nekoreliranih slučajnih varijabli oblika (2.3). Stoga ćemo pretpostaviti da za $i = 1, \ldots, n$ vrijedi

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i, \qquad Var(Y_i) = \sigma^2, \qquad Cov(Y_i, Y_j) = 0,\ i \neq j.$$

Podatke promatramo kao uređene parove $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Ako opaženu vrijednost slučajne varijable Y_i zapišemo kao $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ tako da ϵ_i predstavlja razliku između stvarno opažene vrijednosti u i-tom ponavljanju pokusa i teorijske vrijednosti $\mathrm{E}(Y_i)$. Sada želimo povući pravac kroz $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ tako da se na neki način minimizira udaljenost tih točaka od pravca, tj. biramo pravac koji minimizira neku funkciju od $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$. Različite metode koriste različite funkcije od ϵ_i. Metoda najmanjih kvadrata minimizira sumu kvadrata udaljenosti od pravca, tj. traže se vrijednosti od β_0 i β_1, nazovimo ih $\hat{\beta}_0$ i $\hat{\beta}_1$ koji minimiziraju sumu

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

[3]engleski statističar i antropolog

Parcijalnim deriviranjem funkcije S po β_0 i β_1 te izjednačavanjem parcijalnih derivacija s 0 dobijemo procjenitelje $\hat{\beta}_0$ i $\hat{\beta}_1$ kao rješenja jednadžbi

$$-2\sum_{i=1}^{n}[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] = 0$$

$$-2\sum_{i=1}^{n} x_i[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] = 0.$$

Rješavanjem sustava slijedi

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \\
&= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}, \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}.
\end{aligned}$$

Pravac $y = \hat{\beta}_0 + \hat{\beta}_1 x$ minimizira sumu kvadrata grešaka između opaženih vrijednosti i točaka pravca. Neka je

$$\text{SSE} = \sum_{i=1}^{n} \hat{\epsilon_i}^2 = \sum_{i=1}^{n}[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i]^2.$$

Vrijednosti $\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ nazivamo rezidualima, a SSE (*engl.* error sum of squares) sumom kvadrata reziduala. Metoda najmanjih kvadrata ne daje procjenitelja za σ^2, ali može se pokazati da je nepristran procjenitelj za σ^2 dan s

$$\tilde{\sigma}^2 = \frac{SSE}{n-2}.$$

Zapis $\tilde{\sigma}^2$ koristit ćemo za nepristranog procjenitelja od σ^2. SSE možemo zapisati na sljedeći način:

$$\text{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i.$$

Zapis $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ ćemo koristiti za predikciju vrijednosti Y, a istom vrijednosti procjenjivat ćemo očekivanu vrijednost od Y, tj. procjenitelj za $\text{E}(Y) = \beta_0 + \beta_1 x$ je dan s

$$\hat{E}(Y) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Procjenitelj najmanjih kvadrata je linearna funkcija od $Y_1, Y_2, \ldots, Y_n$ i može se pokazati kako je on procjenitelj s najmanjom varijancom među ostalim linearnim nepristranim procjeniteljima. Stoga se on često naziva linearnim nepristranim procjeniteljem s najmanjom varijancom (*engl.* best linear unbiased estimator).

Teorem 2.1. *Ako je* $E(Y_i) = \beta_0 + \beta_1 x_i$, $Var(Y_i) = \sigma^2$ *i* $Cov(Y_i, Y_j) = 0$ *za* $i \neq j$ *i* $i = 1, \ldots, n$, *tada procjenitelj najmanjih kvadrata ima sljedeća svojstva:*

1. $E(\hat{\beta}_1) = \beta_1$, $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

2. $E(\hat{\beta}_0) = \beta_0$, $Var(\hat{\beta}_0) = \dfrac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

3. $E(c_1\hat{\beta}_0 + c_2\hat{\beta}_1) = c_1\beta_0 + c_2\beta_1$

4. $c_1\hat{\beta}_0 + c_2\hat{\beta}_1$ *je najbolji linearni nepristran procjenitelj za* $c_1\beta_0 + c_2\beta_1$.

Dokaz se može pronaći u [1].

2.2 Višestruka linearna regresija

Promatramo model linearne regresije oblika (2.2) te pretpostavljamo da je realizacija y_i ovisne slučajne varijable Y zabilježena pri vrijednostima $x_{i0}, x_{i1}, \ldots, x_{ip}$, $\quad i = 1, \ldots, n$ tako da je $n \geq p + 1$. Pretpostavljamo da vrijedi

$$\mathrm{E}(Y_i) = \sum_{j=0}^{p} \beta_j x_{ij} \qquad Var(Y_i) = \sigma^2 \qquad Cov(Y_i, Y_j) = 0, \quad i \neq j.$$

Označimo s V matricu kojoj je element u i-tom retku i j-tom stupcu kovarijanca varijabli Y_i i Y_j, tj. $V = \{Cov(Y_i, Y_j)\}$. Matrica V naziva se matrica kovarijanci od $Y_1, Y_2, \ldots, Y_n$. Definirat ćemo očekivanje vektora slučajnih varijabli kao vektor očekivanja pripadajućih slučajnih varijabli. Model višestruke linearne regresije može se zapisati u matričnom obliku na sljedeći način:

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \qquad V = \sigma^2 I,$$

gdje je I jedinična matrica, a $\mathbf{Y}, \boldsymbol{\beta}$ i $\mathbf{X}$ su

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} x_{10} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n0} & \ldots & x_{np} \end{bmatrix}.$$

Procjenitelji slijede iz minimizacije funkcije

$$S = \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right]^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Pristup korišten kod jednostavne linearne regresije lako se generalizira. Izjednačimo parcijalne derivacije od S s 0.

$$\frac{\partial S}{\partial \beta_k} = \sum_{i=1}^{n} 2\Big[y_i - \sum_{j=0}^{p} \beta_j x_{ij}\Big](-x_{ik}) = 0 \qquad k = 0, 1, \ldots, p.$$

Taj sustav je linearan u parametrima i može se zapisati u obliku sljedeće matrične jednadžbe:

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}.$$

Ukoliko je matrica $\mathbf{X}^T\mathbf{X}$ nesingularna, postoji jedinstveno rješenje oblika

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Osim ako je drukčije naglašeno, pretpostavit ćemo da je $\mathbf{X}^T\mathbf{X}$ nesingularna.

Procjenitelji $\hat{\beta}_j$ linearne su funkcije od $Y_1, Y_2, \ldots, Y_n$ i može se pokazati kako su oni nepristrani procjenitelji.

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}(\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

Kao i kod modela jednostavne linearne regresije, procjenitelji najmanjih kvadrata za β_j-ove često se nazivaju linearnim nepristranim procjeniteljima s najmanjom varijancom.

Također se može pokazati da su varijance i kovarijance nepristranog linearnog procjenitelja s najmanjom varijancom elementi matrice $\boldsymbol{C} = \{Cov(\hat{\beta}_i, \hat{\beta}_j)\} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ i da je linearan nepristran procjenitelj s najmanjom varijancom bilo koja linearna kombinacija β_j-ova, recimo $\boldsymbol{r}^T\hat{\boldsymbol{\beta}} = \sum_{j=0}^{p} r_j\hat{\beta}_j$ je linearan nepristran procjenitelj s najmanjom varijancom za $\boldsymbol{r}^T\boldsymbol{\beta} = \sum_{j=0}^{p} r_j\beta_j$.

2.3 Eksponencijalna familija distribucija i generalizirani linearni modeli

Kod modela logističke regresije ovisna varijabla nije modelirana neprekidnom slučajnom varijablom, nego je modelirana binomnom slučajnom varijablom čija distribucija pripada eksponencijalnoj familiji. Model logističke regresije pripada klasi generaliziranih linearnih modela. Stoga ćemo se prvo upoznati s pojmovima eksponencijalne familije distribucija i generaliziranim linearnim modelima.

Definicija 2.2. *Neka je Y slučajna varijabla čija vjerojatnosna distribucija ovisi o parametru θ. Distribucija slučajne varijable Y pripada* ***eksponencijalnoj familiji*** *ako se njena funkcija gustoće može zapisati u sljedećem obliku:*

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)}, \tag{2.4}$$

gdje su a, b, s i t poznate funkcije (vidi [7]).

Ako definiramo $s(y)$ kao $s(y) = e^{d(y)}$, a $t(\theta)$ kao $t(\theta) = e^{c(\theta)}$, možemo uočiti simetriju između y i θ u funkciji (2.4). Tada funkcija (2.4) poprima sljedeći oblik

$$f(y;\theta) = \exp\big[a(y)b(\theta) + c(\theta) + d(y)\big]. \tag{2.5}$$

Ako je $a(y) = y$, kažemo da distribucija ima kanonsku (standardnu) formu i $b(\theta)$ nazivamo prirodnim parametrom distribucije. Ako uz parametar θ postoje drugi parametri, njih smatramo poznatima i ne procjenjujemo ih. Eksponencijalnoj familiji distribucija pripadaju Poissonova, binomna i normalna distribucija.

Funkcija gustoće Bernoullijeve distribucije je oblika

$$f(y) = \pi^y(1-\pi)^{1-y}, \qquad y \in \{0,1\}, \tag{2.6}$$

gdje $\pi = \mathrm{P}(y = 1)$ označava vjerojatnost uspjeha.

Slučajna varijabla Y koja broji uspješne realizacije $n \in \mathbb{N}$ nezavisnih ponavljanja Bernoullijevog pokusa s istom vjerojatnosti uspjeha π naziva se binomna slučajna varijabla. Njena funkcija gustoće dana je s

$$h(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \qquad y \in \{0,1,\ldots,n\}. \tag{2.7}$$

Pokažimo da binomna distribucija pripada eksponencijalnoj familiji distribucija.

$$\begin{aligned}\binom{n}{y}\pi^y(1-\pi)^{n-y} &= \exp\left[\ln\binom{n}{y} + y\ln\pi + (n-y)\ln(1-\pi)\right] \\ &= \exp\left[\ln\frac{n!}{y!(n-y)!} + y(\ln\pi - \ln(1-\pi)) + n\ln(1-\pi)\right] \\ &= \exp\left[y\ln\frac{\pi}{1-\pi} + n\ln(1-\pi) + \ln\frac{n!}{y!(n-y)!}\right]\end{aligned} \tag{2.8}$$

Binomna distribucija pripada eksponencijalnoj familiji distribucija s poznatim funkcijama:

$$\begin{aligned} a(y) &= y & b(\pi) &= \ln\frac{\pi}{1-\pi} \\ c(\pi) &= n\ln(1-\pi) & d(y) &= \ln\frac{n!}{y!(n-y)!}. \end{aligned}$$

Uvrštavanjem $n = 1$ i $y \in \{0,1\}$ u (2.8) pokazujemo kako i Bernoullijeva distribucija pripada eksponencijalnoj familiji. Tada vrijedi: $a(y) = y$, $b(\pi) = \ln\frac{\pi}{1-\pi}$, $d(y) = 0$ i $c(\pi) = ln(1-\pi)$.

Generalizirani linearni model definiran je u terminima skupa međusobno nezavisnih slučajnih varijabli $Y_1, \ldots, Y_N$ čije distribucije ne moraju nužno biti jednake, ali pripadaju eksponencijalnoj familiji i zadovoljavaju sljedeća svojstva (vidi [7]):

1. Distribucija svakog Y_i ima kanonsku formu i ovisi o jednom parametru θ_i (θ_i-evi ne moraju biti jednaki), tj.

$$f_{Y_i}(y;\theta_i) = \exp\big[yb_i(\theta_i) + c_i(\theta_i) + d_i(y)\big], \qquad i = 1,\ldots,N.$$

2. Distribucije svih Y_i istog su tipa pa funkcije b, c i d ne ovise o indeksu i. Zajednička funkcija gustoće slučajnih varijabli $Y_1, \ldots, Y_N$ tada ima sljedeći oblik:

$$\begin{aligned} f(y_1, \ldots, y_N; \theta_1, \ldots, \theta_N) &= \prod_{i=1}^{N} \exp\left[y_i b(\theta_i) + c(\theta_i) + d(y_i)\right] \\ &= \exp\left[\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right]. \end{aligned} \tag{2.9}$$

Skup parametara $\theta_1, \theta_2, \ldots, \theta_N$ inače nije od važnosti za specifikaciju modela pa promatramo manji skup parametara $\beta_1, \beta_2, \ldots, \beta_p$, $(p < N)$. Pretpostavimo da je $\mathrm{E}(Y_i) = \mu_i$, pri čemu je μ_i funkcija od θ_i. Za generalizirani linearni model postoji transformacija od μ_i takva da vrijedi

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

U toj jednadžbi:

- g je monotona, diferencijabilna funkcija i naziva se link funkcija
- vektor $\mathbf{x}_i$ je p-dimenzionalni vektor opisnih varijabli, tj. $\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$
- $\boldsymbol{\beta}$ je p-dimenzionalni vektor parametara, $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$.

Vektor $\mathbf{x}_i^T$ je i-ti redak matrice dizajna $\mathbf{X}$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & \ldots & x_{Np} \end{bmatrix}.$$

Prema [7] generalizirani linearni modeli definirani su kroz tri komponente:

1. Zavisne varijable $Y_1, \ldots, Y_N$ koje imaju isti tip distribucije iz eksponencijalne familije distribucija.
2. Skup parametara $\boldsymbol{\beta}$ i nezavisne varijable čije mjerene vrijednosti kreiraju matricu dizajna $\mathbf{X}$.
3. Monotona link funkcija g takva da je $\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$, a $\mu_i = \mathrm{E}(Y_i)$.

Za binomnu distribuciju vrijedi da je $\mu_i \in (0, 1)$ i za link funkciju bi trebalo vrijediti: $g : (0, 1) \to \mathbb{R}$. Najčešće korištena link funkcija je *logit* funkcija $g(\mu_i) = \log \frac{\mu_i}{1-\mu_i}$.

2.4 Model logističke regresije

Neka je dano n mjerenja zavisne varijable i prediktora $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. Organiziramo dane podatke u dvije matrice:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ x_{21} & \ldots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}$$

Matrica $\mathbf{y}$ predstavlja podatke ovisne varijable, a $\mathbf{x}$ podatke prediktora. Formiramo slučajni uzorak $(Y_1, Y_2, \ldots, Y_n)$ za koji $\mathbf{y}$ čini jednu realizaciju.

Promatramo nelinearan regresijski model u kojem je zavisna varijabla modelirana Bernoullijevom distribucijom

$$Y_i \sim \begin{pmatrix} 0 & 1 \\ 1-\pi_i & \pi_i \end{pmatrix},$$

s parametrom $\mathrm{E}[Y_i] = \pi_i$. Za binomnu distribuciju vrijedi da je $\pi_i \in (0,1)$ i za link funkciju treba vrijediti: $g : (0,1) \to \mathbb{R}$. Umjesto promatranja vjerojatnosti π_i promatramo omjer uspjeha i neuspjeha, tj. šansu $\dfrac{\pi_i}{1-\pi_i}$. Omjer $\dfrac{\pi_i}{1-\pi_i}$ je ograničen na interval $(0,1)$. Logaritam omjera naziva se *logit*:

$$\eta_i = \mathrm{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}). \tag{2.10}$$

Kada vjerojatnost π_i teži prema 0, omjer teži prema 0, a logit prema $-\infty$. S druge strane, kada vjerojatnost π_i teži prema 1, omjer teži prema 1, a logit prema ∞. Logit preslikava interval $(0,1)$ na $\mathbb{R}$. Transformiranjem (2.10) vrijedi:

$$\pi_i = \mathrm{logit}^{-1}(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}.$$

Promatramo model dihotomne logističke regresije

$$\mathrm{logit}(\pi_i) = \mathbf{x_i}^T\boldsymbol{\beta}, \tag{2.11}$$

gdje je $\mathbf{x_i^T} = [x_{i1}, \ldots, x_{ip}]$, a $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ vektor parametara. Djelovanjem eksponencijalne funkcije na jednadžbu (2.11) slijedi da je i-ti omjer dan s

$$\frac{\pi_i}{1-\pi_i} = \exp(\mathbf{x_i}^T\boldsymbol{\beta}). \tag{2.12}$$

Jedan od problema logističke regresije je interpretacija parametara. Na primjer, povećanje j-tog prediktora za jedan, uz sve ostale varijable nepromijenjene povećava omjer vjerojatnosti u smislu množenja s e^{β_j}. S druge strane, iz (2.12) vrijedi:

$$\pi_i = \frac{\exp(\mathbf{x_i}^T\boldsymbol{\beta})}{1+\exp(\mathbf{x_i}^T\boldsymbol{\beta})}. \tag{2.13}$$

Učinak povećanja jednog prediktora na π_i, uz ostale varijable nepromijenjene, ne možemo izraziti na jednostavan način jer je funkcija s desne strane prethodne jednakosti nelinearna funkcija prediktora.

Parametre modela logističke regresije procijenit ćemo metodom maksimalne vjerodostojnosti. Prvo ćemo definirati funkciju vjerodostojnosti.

Definicija 2.3. *Za danu realizaciju* $\mathbf{x} = (x_1, \ldots, x_n)$ *slučajnog uzorka* $(X_1, \ldots, X_n)$ *iz gustoće* $f(\mathbf{x}; \boldsymbol{\theta})$ ***funkcija vjerodostojnosti*** *je*

$$L(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}),$$

gustoća za fiksni $\mathbf{x}$ *promatrana kao funkcija parametra.(vidi [2]).*

Metoda maksimalne vjerodostojnosti temelji se na ideji definiranja procjenitelja maksimizacijom funkcije vjerodostojnosti.

Definicija 2.4. *Neka je* $\mathbf{X}$ *slučajni uzorak iz funkcije gustoće* $f(\mathbf{x}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. *Uz danu realizaciju tog uzorka* $\mathbf{x}$, *neka funkcija vjerodostojnosti* $L(\boldsymbol{\theta}; \mathbf{x})$ *postiže svoj maksimum po* $\boldsymbol{\theta}$ *u* $s(\mathbf{x})$, *tj.*

$$\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x}) = L(s(\mathbf{x}); \mathbf{x}), \qquad s(\mathbf{x}) \in \Theta.$$

Tada statistiku $S = s(\mathbf{X})$ *zovemo* ***procjenitelj maksimalne vjerodostojnosti*** *(engl. Maximum Likelihood Estimator(MLE))(vidi [2]).*

Često je lakše optimizirati logaritam funkcije $L(\boldsymbol{\theta}; \mathbf{x})$. Ako je pretpostavka o tipu distribucije ispravna, tj. ako model odgovara postavljenom problemu, dio mogućih realizacija na kojemu je funkcija gustoće jednaka 0 se neće realizirati. Za danu realizaciju slučajnog uzorka $\mathbf{x}$, funkcija vjerodostojnosti je pozitivna pa ju možemo logaritmirati. Logaritmiranjem se neće promijeniti vrijednost parametra u kojem funkcija vjerodostojnosti postiže maksimum jer je logaritam monotona funkcija. Logaritam funkcije vjerodostojnosti označavamo s $l(\boldsymbol{\theta}; \mathbf{x}) = \log L(\boldsymbol{\theta}; \mathbf{x})$. Odredimo sada procjenitelja maksimalne vjerodostojnosti za model logističke regresije.

Kako slučajnu varijablu Y_i modeliramo Bernoullijevom slučajnom varijablom, njena funkcija gustoće dana je s

$$f_i(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}, \qquad Y_i \in \{0, 1\}, \qquad i = 1, \ldots, n.$$

Kako su $Y_1, Y_2, \ldots, Y_n$ međusobno nezavisne, njihova zajednička funkcija gustoće je sljedećeg oblika:

$$g(Y_1, \ldots, Y_n; \pi_1, \ldots, \pi_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}.$$

Kako bismo lakše pronašli procjenitelja maksimalne vjerodostojnosti, dalje ćemo raditi s logaritmom zajedničke funkcije gustoće.

$$\begin{aligned} \log g(Y_1, \ldots, Y_n; \pi_1, \ldots, \pi_n) &= \log \Big[\prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} \Big] \\ &= \sum_{i=1}^{n} \big[Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i) \big] \\ &= \sum_{i=1}^{n} \left[Y_i \log \frac{\pi_i}{1 - \pi_i} \right] + \sum_{i=1}^{n} \log(1 - \pi_i) \end{aligned} \tag{2.14}$$

Podsjetimo se da vrijedi

$$\log(\frac{\pi_i}{1-\pi_i}) = \mathbf{x_i}^T\boldsymbol{\beta}.$$

Kako je $\mathrm{E}[Y_i] = \pi_i$ za Bernoullijevu slučajnu varijablu, iz (2.13) slijedi da je

$$1 - \pi_i = [1 + \exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]^{-1}.$$

Sada (2.14) možemo zapisati u sljedećem obliku:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(\boldsymbol{x_i}^T\boldsymbol{\beta}) - \sum_{i=1}^{n} \log[1 + \exp(\boldsymbol{x_i}^T\boldsymbol{\beta})]. \tag{2.15}$$

Zamjenjujemo zapis $g(Y_1, \ldots, Y_n; \pi_1, \ldots, \pi_n)$ s $L(\boldsymbol{\beta})$ kako bismo naglasili da sada gledamo logaritam funkcije vjerodostojnosti parametara koje procjenjujemo uz dane podatke. Vrijednosti $\beta_1, \ldots, \beta_p$ koje maksimiziraju logaritam funkcije vjerodostojnosti računaju se numeričkim metodama.

2.5 Model logističke regresije primijenjen u istraživanju i način odabira modela

Za kreiranje modela logističke regresije korištena je funkcija glm() iz paketa stats besplatnog softverskog okruženja R. Funkcija glm() koristi se za usklađivanje generaliziranog linearnog modela. Modele kreirane pomoću glm() funkcije uspoređujemo pomoću Anova testa i Akaike informacijskog kriterija.

2.5.1 Anova

Model bez nezavisnih varijabli nazivamo nul modelom jer ga navodimo u nul hipotezi. Dodajući k nezavisnih varijabli za kreiranje punog modela možemo istražiti doprinose li dodane nezavisne varijable kvaliteti modela.

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \beta_1 \neq 0, \beta_2 \neq 0, \ldots, \beta_k \neq 0 \end{aligned} \tag{2.16}$$

Hipoteze se mogu interpretirati na sljedeći način:

H_0 : Nezavisna varijabla x_i, $\forall i = 1, \ldots, k$ ne utječe na zavisnu varijablu Y

H_1 : Nezavisna varijabla x_i, $i = 1, \ldots, k$ utječe na zavisnu varijablu Y.

2.5.2 AIC

Akaike informacijski kriterij (*engl.* Akaike Information Criterion) je još jedna mjera dobrote prilagodbe modela temeljena na logaritmu funkcije vjerodostojnosti modificirana uvrštavanjem broja parametara tako da penalizira modele s većim brojem parametara. Statistika je dana s

$$AIC = -2\ln L + 2p$$

gdje je L maksimum funkcije vjerodostojnosti promatranog modela, a p broj procijenjenih parametara modela. Tu definiciju AIC-a koristi i statistički software R (za više vidi [7]). Modele s manjim AIC-om smatramo boljima od modela s većim AIC-om.

3 Analiza zavisnosti

Ovisnost diskretne zavisne varijable o nezavisnoj varijabli testirat ćemo χ^2 test ako je nezavisna varijabla diskretna, odnosno t-testom ako je nezavisna varijabla neprekidna.

3.1 Pearsonov χ^2 test nezavisnosti za diskretne slučajne varijable

Neka je f_{ij} opažena frekvencija podataka koji pripadaju i-toj kategoriji od X i j-toj kategoriji od Y. Neka su e_{ij} pripadajuće očekivane frekvencije ako su X i Y nezavisne. Pearsonova χ^2 test statistika dana je s

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}.$$

Nul hipotezu o nezavisnosti varijabli odbacujemo ako je p-vrijednost Pearsonove χ^2 test statistike manja od razine značajnosti α (za više vidi [3]).

3.2 t-test za testiranje jednakosti očekivanja

Kada imamo velike uzorke, hipotezu o jednakosti očekivanja slučajnih varijabli možemo testirati neovisno o distribuciji tih varijabli. Neka su n i m veličine uzoraka, $\bar{X_n}$ i $\bar{X_m}$ redom aritmetičke sredine, a σ_n i σ_m redom poznate standardne devijacije tih uzoraka.
Nul-hipoteza: $H_0 : \quad \mu_1 = \mu_2$
Test-statistika:

$$\hat{Z} = \frac{\bar{X_n} - \bar{X_m}}{\sqrt{\frac{\sigma_n}{n} + \frac{\sigma_m}{m}}}$$

U slučaju istinitosti nul-hipoteze, test-statistika $\hat{Z}$ ima približno normalnu distribuciju. Ako sa Z označimo slučajnu varijablu s $\mathcal{N}(0,1)$ distribucijom, na osnovi realizacije $\hat{z}$ statistike $\hat{Z}$ na podacima možemo odrediti p-vrijednost kao $p = P\{Z \geq \hat{z}\}$ ako je alternativna hipoteza oblika $H_1 : \mu_1 - \mu_2 > 0$. Odnosno $p = P\{Z \leq \hat{z}\}$ za alternativnu hipotezu oblika $H_1 : \mu_1 - \mu_2 < 0$. Ako je tako izračunata p-vrijednost $< \alpha$, odbacujemo nul-hipotezu na razini značajnosti α i prihvaćamo alternativnu hipotezu. S druge strane, ako je $p > \alpha$, zaključujemo da nemamo dovoljno argumenata za odbacivanje nul-hipoteze (vidi [3]). Aritmetičke sredine uzoraka X_n i X_m koristimo kao procjenitelje za očekivanja μ_1 i μ_2. Za primjenu ovog testa potrebno je poznavati i varijance σ_n^2 i σ_m^2, što u primjeni najčešće nije slučaj. Međutim, u slučaju velikih uzoraka možemo iskoristiti korigirane varijance uzoraka s_n^2 i s_m^2 kao procjene nepoznatih varijanci.

4 Empirijsko istraživanje: Modeliranje odljeva igrača u online klađenjima

4.1 Prethodna istraživanja

Coussement i De Bock u [6] predviđaju odljev klijenata kladionice Bwin, online priređivača igara na sreću. Zaključili su kako algoritmi učenja ansamblom (*engl.* ensemble learning)[4] regresijskih stabala i generaliziranih aditivnih modela postižu bolje rezultate od jednostavnih metoda poput običnih regresijskih stabala i generaliziranih aditivnih modela (kao što je logistička regresija). Metode učenja ansamblom zahtijevaju posebna znanja radi razumijevanja algoritma i u prosjeku su vremenski zahtjevnije od jednostavnih metoda. Vrijeme zadnjeg posjeta, učestalost posjeta i gubitak, odnosno dobitak novca pri zadnjem posjetu pokazale su se kao najznačajnije varijable u modeliranju.

Analiza odljeva klijenata puno je učestalija u području telekomunikacija jer pružatelji usluga uočavaju gubitke zbog prelaska klijenata drugom pružatelju usluga. U svom radu [8], Huang, Kechadi i Buckley predviđali su odljev klijenata pomoću sedam različitih metoda među kojima su logistička regresija, stabla odlučivanja, neuronske mreže i stroj potpornih vektora (*engl.* Support Vector Machines(SVM)). Pokazalo se da ako za kriterij gledamo stvarnu stopu odljeva i lažnu stopu odljeva, stabla odlučivanja i SVM daju najbolje rezultate. Ako nas pak zanima vjerojatnost odljeva pojedinog klijenta, najbolje je modelirati logističkom regresijom. Najznačajnijim varijablama za modeliranje su se pokazale informacije o računima, vrste poziva, detalji poziva, Henleyeva segmentacija[5], informacije o korisniku i način plaćanja.

Ismail i drugi u radu [9] modelirali su odljev klijenata logističkom regresijom i neuronskom mrežom. Najboljim se pokazao model neuronske mreže s 14 ulaznih podataka, jednim skrivenim slojem i jednim izlaznim slojem. Taj model ostvario je točnost predikcije od $91,28\%$. Modeliranje odljeva neuronskom mrežom pokazao se kao dobra alternativa tradicionalnim predikcijskim modelima poput logističke regresije.

4.2 Varijable i podaci

U ovom radu za provođenje istraživanja koristili smo bazu podataka jednog online priređivača igara na sreću. U bazi se nalaze podaci o 1200 klijenata od kojih je njih 600 označeno kao aktivni klijenti, a 600 kao neaktivni klijenti. Za kreiranje modela korišteno je 70% ukupnih podataka, tj. 420 aktivnih i 420 neaktivnih klijenata. Preostalih 30% podataka (od kojih 180 aktivnih i 180 neaktivnih) korišteno je za validaciju modela.

Zavisna varijabla churn dijeli klijente na aktivne (označene 0) i neaktivne (označene 1). Aktivni klijenti su oni koji sudjeluju u online igrama na sreću, odnosno oni koji se klade, a neaktivni klijenti su oni koji su značajno smanjili ili u potpunosti prestali s klađenjem. Osim zavisne varijable churn, baza sadrži 41 nezavisnu varijablu kojima je okarakteriziran svaki od 1200 klijenata.

- Age - numerička varijabla koja opisuje starosnu dob klijenta

[4] Ansambli su modeli strojnog učenja temeljeni na višestrukim osnovnim modelima.

[5] Henleyeva segmentacija svrstava klijente u grupe s obzirom na njihove karakteristike, potrebe i njihovu komercijalnu vrijednost. Za više vidi [8].

- PayIn_ATG - numerička varijabla koja opisuje iznos uplaćen u jednom mjesecu za ATG[6]
- PayIn_INTERNETDOGS - numerička varijabla koja opisuje iznos uplaćen u jednom mjesecu za utrke pasa
- PayIn_INTERNETLIVE - numerička varijabla koja opisuje iznos uplaćen u jednom mjesecu za klađenje uživo
- PayIn_LOTTO - numerička varijabla koja opisuje iznos uplaćen u jednom mjesecu za klađenje na izvlačenje loto brojeva u svijetu, osim na rezultate izvlačenja lota Hrvatske Lutrije
- PayIn_SPORT - numerička varijabla koja opisuje iznos uplaćen u jednom mjesecu za klađenje na sportske događaje
- Win_ATG - numerička varijabla koja opisuje iznos koji je igrač osvojio u jednom mjesecu za klađenje na ATG
- Win_INTERNETDOGS - numerička varijabla koja opisuje iznos koji je igrač osvojio u jednom mjesecu za klađenje na utrke pasa
- Win_INTERNETLIVE - numerička varijabla koja opisuje iznos koji je igrač osvojio u jednom mjesecu za klađenje uživo
- Win_LOTTO - numerička varijabla koja opisuje iznos koji je igrač osvojio u jednom mjesecu za klađenje na izvlačenje loto brojeva u svijetu, osim na rezultate izvlačenja lota Hrvatske Lutrije
- Win_SPORT - numerička varijabla koja opisuje iznos koji je igrač osvojio u jednom mjesecu za klađenje na sportske događaje
- Lost_ATG - numerička varijabla koja opisuje iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na ATG
- Lost_INTERNETDOGS - numerička varijabla koja opisuje iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na utrke pasa
- Lost_INTERNETLIVE - numerička varijabla koja opisuje iznos koji je igrač izgubio u jednom mjesecu zbog klađenja uživo
- Lost_LOTTO - numerička varijabla koja opisuje iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na izvlačenje loto brojeva u svijetu, osim na rezultate izvlačenja lota Hrvatske Lutrije
- Lost_SPORT - numerička varijabla koja opisuje iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na sportske događaje

[6]ATG je trgovačko društvo, osnovano od strane Švedskog kasačkog i galopskog saveza s licencom priređivanja klađena na konjičke utrke u Švedskoj.

- BetCount_ATG - numerička varijabla koja opisuje broj oklada na ATG u jednom mjesecu
- BetCount_INTERNETDOGS - numerička varijabla koja opisuje broj oklada na utrke pasa u jednom mjesecu
- BetCount_INTERNETLIVE - numerička varijabla koja opisuje broj oklada za klađenje uživo u jednom mjesecu
- BetCount_LOTTO - numerička varijabla koja opisuje broj loto uplata u jednom mjesecu
- BetCount_SPORT - numerička varijabla koja opisuje broj oklada na sportske događaje u jednom mjesecu
- FixedBetCount_ATG - numerička varijabla koja opisuje broj fiksnih oklada na ATG u jednom mjesecu
- FixedBetCount_INTERNETDOGS - numerička varijabla koja opisuje broj fiksnih oklada na utrke pasa u jednom mjesecu
- FixedBetCount_INTERNETLIVE - numerička varijabla koja opisuje broj fiksnih oklada za klađenje uživo u jednom mjesecu
- FixedBetCount_LOTTO - numerička varijabla koja opisuje broj fiksnih loto uplata u jednom mjesecu
- FixedBetCount_SPORT - numerička varijabla koja opisuje broj fiksnih oklada na sportske događaje u jednom mjesecu
- UnfixedBetCount_ATG - numerička varijabla koja opisuje broj nefiksnih oklada na ATG u jednom mjesecu
- UnfixedBetCount_INTERNETDOGS - numerička varijabla koja opisuje broj nefiksnih oklada na utrke pasa u jednom mjesecu
- UnfixedBetCount_INTERNETLIVE - numerička varijabla koja opisuje broj nefiksnih oklada za klađenje uživo u jednom mjesecu
- UnfixedBetCount_LOTTO - numerička varijabla koja opisuje broj nefiksnih loto uplata u jednom mjesecu
- UnfixedBetCount_SPORT - numerička varijabla koja opisuje broj nefiksnih oklada na sportske događaje u jednom mjesecu
- ProsloDanaOdZadnjegGubitka - numerička varijabla koja opisuje koliko je dana prošlo od posljednjeg gubitka
- ProsloDanaOdZadnjegDobitka - numerička varijabla koja opisuje koliko je dana prošlo od posljednjeg dobitka
- ProsloDanaOdZadnjeAkcije - numerička varijabla koja opisuje koliko je dana prošlo od posljednjeg klađenja

- BrojDanaOdregistracijeDoZadnjeAktivnosti - numerička varijabla koja opisuje koliko je dana prošlo od registracije do zadnjeg klađenja
- ProsjecnoDanaIzmedjuKladjenja - numerička varijabla koja opisuje prosječan broj dana koji prođe između dva klađenja
- MinimalanBrojDanaIzmedjuKladjenja - numerička varijabla koja opisuje minimalan broj dana koji prođe između dva klađenja
- MaksimalanBrojDanaIzmedjuKladjenja - numerička varijabla koja opisuje maksimalan broj dana koji prođe između dva klađenja
- BrojKladjenjaUZadnjemTjednu - numerička varijabla koja opisuje broj klađenja u zadnjem tjednu
- BrojKladjenjaUZadnjemMjesecu - numerička varijabla koja opisuje broj klađenja u zadnjem mjesecu
- Bonus - kategorijalna varijabla koja poprima vrijednost 1 ako je igrač dobio bonus prilikom registracije i 0 ako nije.

Nedostajuće vrijednosti pojavljuju se u varijablama ProsloDanaOdZadnjegGubitka, ProsloDanaOdZadnjegDobitka i ProsloDanaOdZadnjeAkcije pa te varijable pretvaramo u kategorijalne. Podijelit ćemo varijable tako da u svim grupama bude približno jednako klijenata, a zatim ćemo klijente s nedostajućim vrijednostima pridružiti onoj grupi koja ima najsličniji omjer aktivnih i neaktivnih klijenata.

Prošlo dana od zadnjeg gubitka	**Broj neaktivnih**	**Broj aktivnih**
< 7	27	246
≥ 7 i < 36	114	189
≥ 36 i < 78	178	104
≥ 78	243	35
nedostajuća vrijednost	38	26

Tablica 2: Frekvencije varijable ProsloDanaOdZadnjegGubitka s obzirom na varijablu churn

U tablicama 2, 3 i 4 dane su frekvencije varijabli ProsloDanaOdZadnjegGubitka, ProsloDanaOdZadnjegDobitka i ProsloDanaOdZadnjeAkcije podijeljenih po grupama s obzirom na varijablu churn. S obzirom na grupirane podatke varijable ProsloDanaOdZadnjegGubitka, klijente za koje nemamo podatak priključujemo grupi klijenata kojima je od posljednjeg gubitka prošlo barem 36, a manje od 78 dana. Slično tome, klijente za koje nemamo podatak o varijabli ProsloDanaOdZadnjegDobitka priključujemo grupi klijenata kojima je od posljednjeg dobitka prošlo barem 60, a manje od 93 dana. Klijente za koje nemamo podatak o varijabli ProsloDanaOdZadnjeAkcije priključujemo grupi klijenata kojima je od posljednje akcije prošlo barem 45, a manje od 78 dana. Tako kreirane varijable nazvat ćemo odzadnjeggubitka, odzadnjegdobitka i odzadnjeakcije.

Prošlo dana od zadnjeg dobitka	**Broj neaktivnih**	**Broj aktivnih**
< 19	42	238
≥ 19 i < 60	149	178
≥ 60 i < 93	141	100
≥ 93	230	58
nedostajuća vrijednost	38	26

Tablica 3: Frekvencije varijable ProsloDanaOdZadnjegDobitka s obzirom na varijablu churn

Prošlo dana od zadnje akcije	**Broj neaktivnih**	**Broj aktivnih**
< 10	36	287
≥ 10 i < 45	149	173
≥ 45 i < 78	134	80
≥ 78	243	34
nedostajuća vrijednost	38	26

Tablica 4: Frekvencije varijable ProsloDanaOdZadnjeAkcije s obzirom na varijablu churn

Na skupu podataka za treniranje testiramo zavisnost varijable churn o nezavisnim varijablama kako je opisano u poglavlju 3. P-vrijjednosti χ^2 testa o nezavisnosti varijable churn i diskretnih slučajnih varijabli dane su u tablici 5, a u tablici 6 dane su p-vrijednosti t-testa o jednakosti očekivanja promatrane neprekidne varijable s obzirom na varijablu churn. Zbog konstantnosti podataka u nekim varijablama u p-vrijednostima t-testa pojavljuje se NA.

Varijabla	**p-vrijednost**
Bonus	0.2275
od zadnjeg gubitka	$< 2.2e-16$
od zadnjeg dobitka	$< 2.2e-16$
od zadnje akcije	$< 2.2e-16$

Tablica 5: P-vrijednosti χ^2 testa za diskretne slučajne varijable

Varijabla	p-vrijednost
Age	0.5377
PayIn_ATG	0.105
PayIn_INTERNETDOGS	0.1461
PayIn_INTERNETLIVE	0.1574
PayIn_LOTTO	0.3573
PayIn_SPORT	0.5069
Win_ATG	0.2005
Win_INTERNETDOGS	0.155
Win_INTERNETLIVE	0.2046
Win_LOTTO	0.3419
Win_SPORT	0.2703
Lost_ATG	0.0552
Lost_INTERNETDOGS	0.1937
Lost_INTERNETLIVE	0.0012
Lost_LOTTO	0.0052
Lost_SPORT	0.0091
BetCount_ATG	0.2956
BetCount_INTERNETDOGS	0.0181
BetCount_INTERNETLIVE	0.0052
BetCount_LOTTO	0.0079
BetCount_SPORT	1.158e-08
FixedBetCount_ATG	NA

Varijabla	p-vrijednost
FixedBetCount_INTERNETDOGS	0.2168
FixedBetCount_INTERNETLIVE	0.0052
FixedBetCount_LOTTO	NA
FixedBetCount_SPORT	0.0412
UnfixedBetCount_ATG	NA
UnfixedBetCount_INTERNETDOGS	NA
UnfixedBetCount_INTERNETLIVE	NA
UnfixedBetCount_LOTTO	0.0079
UnfixedBetCount_SPORT	3.973e-09
BrojDanaOdregistracije-DoZadnjeAktivnosti	$< 2.2e-16$
ProsjecnoDana IzmedjuKladjenja	0.8547
MinimalanBrojDana IzmedjuKladjenja	NA
MaksimalanBrojDana IzmedjuKladjenja	0.0031
BrojKladjenjaU ZadnjemTjednu	1.184e-06
BrojKladjenjaU ZadnjemMjesecu	4.322e-11

Tablica 6: P-vrijednosti t-testa za neprekidne slučajne varijable

Pomoću varijabli koje su se pokazale značajnima napravimo profil aktivnog i neaktivnog igrača.

1. Iznos koji je igrač izgubio u jednom mjesecu zbog klađenja uživo (Lost_INTERNETLIVE)

   Varijabla Lost_INTERNETLIVE opisuje iznos koji je igrač izgubio u jednom mjesecu zbog klađenja uživo. Minimalan izgubljen iznos je 0,00 kn, a maksimalan izgubljen iznos je 1331,81 kn. U prosjeku je izgubljen iznos od 31,39 kn. U tablici 7 su navedene osnovne numeričke karakteristike s obzirom na aktivnost. Uočavamo kako je u prosjeku aktivan igrač izgubio više novca zbog klađenja uživo od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	0	42,49	25	1331,81
Neaktivni	0	0	0	20,03	5	667,5

Tablica 7: Iznos koji je igrač izgubio u jednom mjesecu zbog klađenja uživo s obzirom na aktivnost

2. Iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na izvlačenje loto brojeva (Lost_LOTTO)

   Varijabla Lost_LOTTO opisuje iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na izvlačenje loto brojeva. Minimalan izgubljen iznos je 0,00 kn, a maksimalan izgubljen iznos je 3671,07 kn. U prosjeku je izgubljen iznos od 29,67 kn. U tablici 8 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač izgubio više novca zbog klađenja na izvlačenje loto brojeva od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	0	45,37	10,63	3671,07
Neaktivni	0	0	0	13,961	4,237	1074,5

Tablica 8: Iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na izvlačenje loto brojeva s obzirom na aktivnost

3. Iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na sportske događaje (Lost_SPORT)

   Varijabla Lost_SPORT opisuje iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na sportske događaje. Minimalan izgubljen iznos je 0,00 kn, a maksimalan izgubljen iznos je 20300,00 kn. U prosjeku je izgubljen iznos od 325,90 kn. U tablici 9 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač izgubio više novca zbog klađenja na sportske događaje od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	76,73	433,96	298,73	20300
Neaktivni	0	0	20	217,8	139,2	8660

Tablica 9: Iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na sportske događaje s obzirom na aktivnost

4. Broj oklada na utrke pasa u jednom mjesecu (BetCount_INTERNETDOGS)

   Varijabla BetCount_INTERNETDOGS opisuje broj oklada na utrke pasa u jednom mjesecu. Minimalan broj oklada je 0, a maksimalan broj oklada je 2919. U prosjeku je broj oklada 19,45. Barem 75% igrača se nije kladilo na utrke pasa. U tablici 10 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj oklada na utrke pasa od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	0	30,8	0	2919
Neaktivni	0	0	0	8,091	0	444

Tablica 10: Broj oklada na utrke pasa u jednom mjesecu s obzirom na aktivnost

5. Broj oklada na klađenje uživo u jednom mjesecu (BetCount_INTERNETLIVE)

   Varijabla BetCount_INTERNETLIVE opisuje broj oklada na klađenje uživo u jednom mjesecu. Minimalan broj oklada je 0, a maksimalan broj oklada je 13530. U prosjeku je broj oklada 128,84. U tablici 11 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj oklada na klađenje uživo od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	9	190,5	92	13530
Neaktivni	0	0	1	67,21	30,25	5823

Tablica 11: Broj oklada na klađenje uživo u jednom mjesecu s obzirom na aktivnost

6. Broj oklada na izvlačenje loto brojeva u jednom mjesecu (BetCount_LOTTO)

   Varijabla BetCount_LOTTO opisuje broj oklada na izvlačenje loto brojeva u jednom mjesecu. Minimalan broj oklada je 0, a maksimalan broj oklada je 9670. U prosjeku je broj oklada 221,8. 50% ili više igrača nije se kladilo na izvlačenje loto brojeva. U tablici 12 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj oklada na izvlačenje loto brojeva od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	5	291,7	131	9670
Neaktivni	0	0	0	152	71	4338

Tablica 12: Broj oklada na izvlačenje loto brojeva u jednom mjesecu s obzirom na aktivnost

7. Broj oklada na sportske događaje u jednom mjesecu (BetCount_SPORT)

   Varijabla BetCount_SPORT opisuje broj oklada na sportske događaje u jednom mjesecu. Minimalan broj oklada je 0, a maksimalan broj oklada je 24232. U prosjeku je broj oklada 716,7. 75% ili više igrača sudjelovalo je u klađenju na sportske događaje. U tablici 13 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj oklada na klađenje na sportske događaje od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	171,2	395,5	1036,2	984,8	24232
Neaktivni	0	106	198	397,2	402,2	10057

Tablica 13: Broj oklada na sportske događaje u jednom mjesecu s obzirom na aktivnost

8. Broj fiksnih oklada na klađenje uživo u jednom mjesecu (FixedBetCount_INTERNETLIVE)

   Varijabla FixedBetCount_INTERNETLIVE opisuje broj fiksnih oklada na klađenje uživo u jednom mjesecu. Minimalan broj fiksnih oklada je 0, a maksimalan broj fiksnih oklada je 13530. U prosjeku je broj fiksnih oklada 128,84. U tablici 14 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj fiksnih oklada na klađenje uživo od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	9	190,5	92	13530
Neaktivni	0	0	1	67,21	30,25	5823

Tablica 14: Broj fiksnih oklada na klađenje uživo u jednom mjesecu s obzirom na aktivnost

9. Broj fiksnih oklada na sportske događaje u jednom mjesecu (FixedBetCount_SPORT)

   Varijabla FixedBetCount_SPORT opisuje broj fiksnih oklada na sportske događaje u jednom mjesecu. Minimalan broj fiksnih oklada je 0, a maksimalan broj fiksnih oklada je 5946. U prosjeku je broj fiksnih oklada 26,24. U tablici 15 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj fiksnih oklada na klađenje na sportske događaje od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	0	42,63	9	5946
Neaktivni	0	0	0	9,843	4	374

Tablica 15: Broj fiksnih oklada na sportske događaje u jednom mjesecu s obzirom na aktivnost

10. Broj nefiksnih oklada na izvlačenje loto brojeva u jednom mjesecu (UnfixedBetCount_LOTTO)

    Varijabla UnfixedBetCount_LOTTO opisuje broj nefiksnih oklada na izvlačenje loto brojeva u jednom mjesecu. Minimalan broj nefiksnih oklada je 0, a maksimalan broj nefiksnih oklada je 9670. U prosjeku je broj nefiksnih oklada 221,8. U tablici 16 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj nefiksnih oklada na izvlačenje loto brojeva od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	5	291,7	131	9670
Neaktivni	0	0	0	152	71	4338

Tablica 16: Broj nefiksnih oklada na izvlačenje loto brojeva u jednom mjesecu s obzirom na aktivnost

11. Broj nefiksnih oklada na sportske događaje u jednom mjesecu (UnfixedBetCount_SPORT)

    Varijabla UnfixedBetCount_SPORT opisuje broj nefiksnih oklada na izvlačenje loto brojeva u jednom mjesecu. Minimalan broj nefiksnih oklada je 0, a maksimalan broj nefiksnih oklada je 22744. U prosjeku je broj nefiksnih oklada 690,5. U tablici 17 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj nefiksnih oklada na klađenje na sportske događaje od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	169	387	993,6	959,8	22744
Neaktivni	0	104,8	193	387,4	388	10057

Tablica 17: Broj nefiksnih oklada na sportske događaje u jednom mjesecu s obzirom na aktivnost

12. Broj dana od registracije do zadnje aktivnosti (BrojDanaOdregistracijeDoZadnjeAktivnosti)

    Varijabla BrojDanaOdregistracijeDoZadnjeAktivnosti opisuje broj dana od registracije do zadnje aktivnosti. Minimalan broj dana je 0, a maksimalan broj dana je 464. U prosjeku je broj dana 182,1. U tablici 18 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj dana od registracije do zadnje aktivnosti od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	82	208	223	376,8	464
Neaktivni	0	40	89	141,2	233	447

Tablica 18: Broj dana od registracije do zadnje aktivnosti s obzirom na aktivnost

13. Maksimalan broj dana koji prođe između dva klađenja (MaksimalanBrojDanaIzmedjuKladjenja)

    Varijabla MaksimalanBrojDanaIzmedjuKladjenja opisuje maksimalan broj dana koji prođe između dva klađenja. Minimalan broj dana je 1, a maksimalan broj dana je 159. U prosjeku je broj dana 22,92. U tablici 19 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći maksimalan broj dana koji prođe između dva klađenja od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	1	10	20	25,08	32	139
Neaktivni	1	6	14	20,76	28	159

Tablica 19: Maksimalan broj dana koji prođe između dva klađenja s obzirom na aktivnost

14. Broj klađenja u zadnjem tjednu (BrojKladjenjaUZadnjemTjednu)

    Varijabla BrojKladjenjaUZadnjemTjednu opisuje broj klađenja u zadnjem tjednu. Minimalan broj klađenja je 0, a maksimalan broj klađenja je 2539. U prosjeku je broj klađenja 36,89. U tablici 20 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj klađenja u zadnjem tjednu od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	0	63,74	31	1767
Neaktivni	0	0	0	10,04	0	2539

Tablica 20: Broj klađenja u zadnjem tjednu s obzirom na aktivnost

15. Broj klađenja u zadnjem mjesecu (BrojKladjenjaUZadnjemMjesecu)

    Varijabla BrojKladjenjaUZadnjemMjesecu opisuje broj klađenja u zadnjem tjednu. Minimalan broj klađenja je 0, a maksimalan broj klađenja je 8367. U prosjeku je broj klađenja 147,2. U tablici 21 navedene su osnovne numeričke karakteristike s obzirom na aktivnost. U prosjeku je aktivan igrač imao veći broj klađenja u zadnjem tjednu od neaktivnog igrača.

	Minimum	Donji kvartil	Medijan	Prosjek	Gornji kvartil	Maksimum
Aktivni	0	0	0	273,6	166	8367
Neaktivni	0	0	0	20,72	0	3545

Tablica 21: Broj klađenja u zadnjem mjesecu s obzirom na aktivnost

Općenito uočavamo da su neaktivni igrači u prosjeku izgubili manje novca na klađenje od aktivnih igrača. U prosjeku neaktivni igrači imaju manji broj oklada od aktivnih igrača, manji broj dana od registracije do posljednje aktivnosti, manji maksimalan broj dana između dva klađenja, manji broj klađenja u zadnjem tjednu i u zadnjem mjesecu. Najviše neaktivnih igrača ima u kategoriji klijenata kojima je od zadnjeg gubitka ili akcije prošlo više od 77 dana, odnosno u kategoriji klijenata kojima je od zadnjeg dobitka prošlo barem 93 dana.

4.3 Model neuronskih mreža za procjenu odljeva igrača u online klađenjima

U našem primjeru, podaci su standardizirani koristeći funkciju StandardScaler iz knjižnice sklearn.preprocessing. Za aktivacijsku funkciju odabrana je logistička funkcija, a za određivanje optimalnih parametara korištena je metoda stohastičkog gradijentnog spusta kao što je opisano u poglavlju 1.3. Sve neuronske mreže trenirane su s jednim skrivenim slojem jer dodavanjem više skrivenih slojeva u našem primjeru mreže postaju pretrenirane. Ulazni sloj svih treniranih neuronskih mreža ima 41 čvor koje čine sve neovisne varijable iz baze podataka pri čemu su varijable ProsloDanaOdZadnjegGubitka, ProsloDanaOdZadnjegDobitka i ProsloDanaOdZadnjeAkcije zamijenjene kategorijalnim varijablama kao što je opisano u poglavlju 4.2. Izlazni sloj treniranih neuronskih mreža sastoji se od dva čvora i predstavlja broj klasa u koje svrstavamo klijente obzirom na ovisnu varijablu churn (aktivni i neaktivni). Na osnovu prethodnih istraživanja za broj čvorova u skrivenom sloju uzima se broj između 2 i 41. U našim primjerima, koristeći funkciju GridSearchCV iz knjižnice sklearn.model_selection za svaki model izabrali smo broj čvorova u skrivenom sloju (15, 20 ili 25 čvorova) za koji model postiže najveću točnost.

Prvi model neuronske mreže MLP1 kreiran je sa 25 čvorova u skrivenom sloju i adaptivnom stopom učenja. Adaptivna stopa učenja ostaje jednaka konstantnoj stopi učenja sve dok se trošak smanjuje. Svaki put kada se u dva uzastopna koraka trošak ne smanji za barem $tol = 0,0001$, stopa učenja u tom trenutku dijeli se s 5. Drugi model neuronske mreže MLP2 kreiran je sa 15 čvorova u skrivenom sloju i konstantnom stopom učenja čija je početna vrijednost zadana s 0,2 i momentum 0. Treći model neuronske mreže MLP3 kreiran je sa 25 čvorova u skrivenom sloju i konstantnom stopom učenja čija je početna vrijednost zadana s 0,2 i momentum 0,9. Tako izrađene modele validiramo na skupu podataka za testiranje.

U tablicama 22, 23 i 24 redom su dane matrice konfuzije modela neuronske mreže MLP1, MLP2 i MLP3.

Matrica konfuzije MLP1

	Predviđeni aktivni	Predviđeni neaktivni
Stvarno aktivni	131	49
Stvarno neaktivni	47	133

Matrica konfuzije MLP1 u postotcima

	Predviđeni aktivni	Predviđeni neaktivni
Stvarno aktivni	36%	14%
Stvarno neaktivni	13%	37%

Tablica 22: Matrica konfuzije MLP1

Analizom težina u slojevima neuronske mreže, varijable s najvećim težinama u svim slojevima u modelu MLP1 su varijabla koja opisuje broj oklada na utrke pasa (BetCount_INTERNETDOGS), broj nefiksnih oklada na sportske događaje (UnfixedBetCount_SPORT), broj dana od zadnjeg gubitka u kategorijama (odzadnjeggubitka), broj dana od zadnjeg dobitka u kategorijama (odzadnjegdobitka), broj dana od zadnje akcije u kategorijama (odzadnjeakcije), broj dana od registracije do zadnje aktivnosti (BrojDanaOdregistracijeDoZadnjeAktivnosti) i broj klađenja u zadnjem mjesecu (BrojKladjenjaUZadnjemMjesecu).

U modelu MLP2 to su iznos izgubljen za klađenje na sportske događaje (Win_SPORT), broj oklada na sportske događaje (BetCount_SPORT), broj dana od zadnjeg gubitka u kategorijama

Matrica konfuzije MLP2

	Predviđeni aktivni	Predviđeni neaktivni
Stvarno aktivni	137	43
Stvarno neaktivni	46	134

Matrica konfuzije MLP2 u postotcima

	Predviđeni aktivni	Predviđeni neaktivni
Stvarno aktivni	38%	12%
Stvarno neaktivni	13%	37%

Tablica 23: Matrica konfuzije MLP2

Matrica konfuzije MLP3

	Predviđeni aktivni	Predviđeni neaktivni
Stvarno aktivni	125	55
Stvarno neaktivni	40	140

Matrica konfuzije MLP3 u postotcima

	Predviđeni aktivni	Predviđeni neaktivni
Stvarno aktivni	35%	15%
Stvarno neaktivni	11%	39%

Tablica 24: Matrica konfuzije MLP3

(odzadnjeggubitka), broj dana od zadnje akcije u kategorijama (odzadnjeakcije), broj dana od registracije do zadnje aktivnosti (BrojDanaOdregistracijeDoZadnjeAktivnosti) i broj klađenja u zadnjem mjesecu (BrojKladjenjaUZadnjemMjesecu).

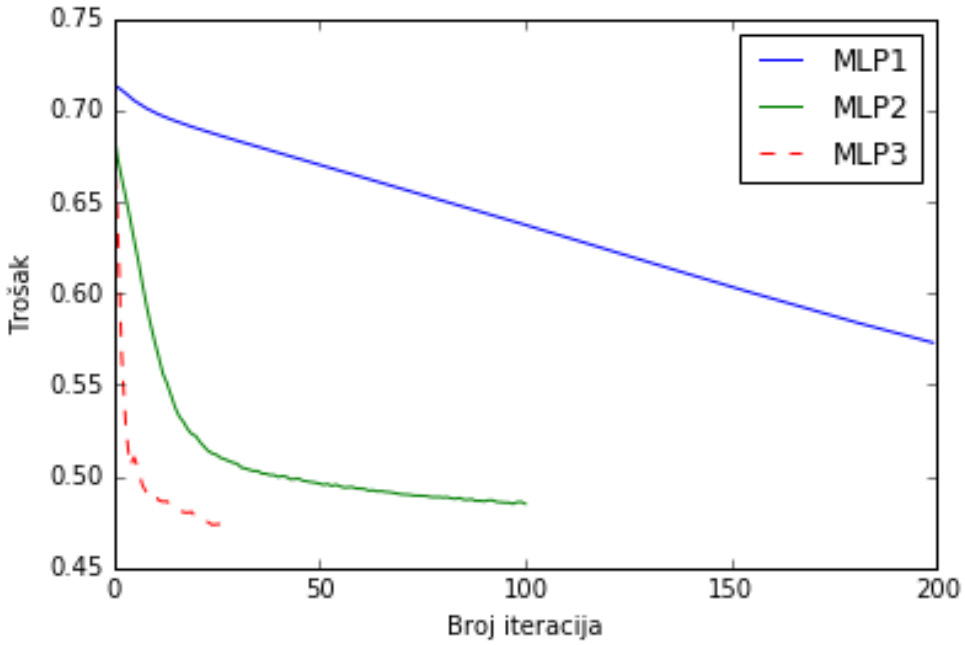
U modelu MLP3 varijable s najvećim težinama su iznos osvojen za klađenje na sportske događaje (Win_SPORT), broj oklada na ATG (BetCount_ATG), broj dana od zadnjeg gubitka u kategorijama (odzadnjeggubitka), prosječan broj dana između dva klađenja (ProsjecnoDanaIzmedjuKladjenja) i maksimalan broj dana između dva klađenja (MaksimalanBrojDanaIzmedjuKladjenja).

Grafove funkcija troška za tri promatrana modela možemo vidjeti na slici 3. Plavom bojom označen je graf funcije troška modela MLP1, zelenom bojom graf funkcije troška modela MLP2, a crvenom bojom graf funkcije troška modela MLP3. Vidimo da funkcija troška modela MLP3 najbrže pada i postiže najmanje vrijednosti. Usporedimo modele s obzirom na postignutu točnost, grešku tipa 1, grešku tipa 2 i AUC-vrijednost u tablici 25. Kako su AUC vrijed-

Model	**Točnost**	**Greška tipa 1**	**Greška tipa 2**	**AUC-vrijednost**
MLP1	0,733	0,26	0,27	0,816
MLP2	0,753	0,26	0,24	0,813
MLP3	0,736	0,22	0,31	0,821

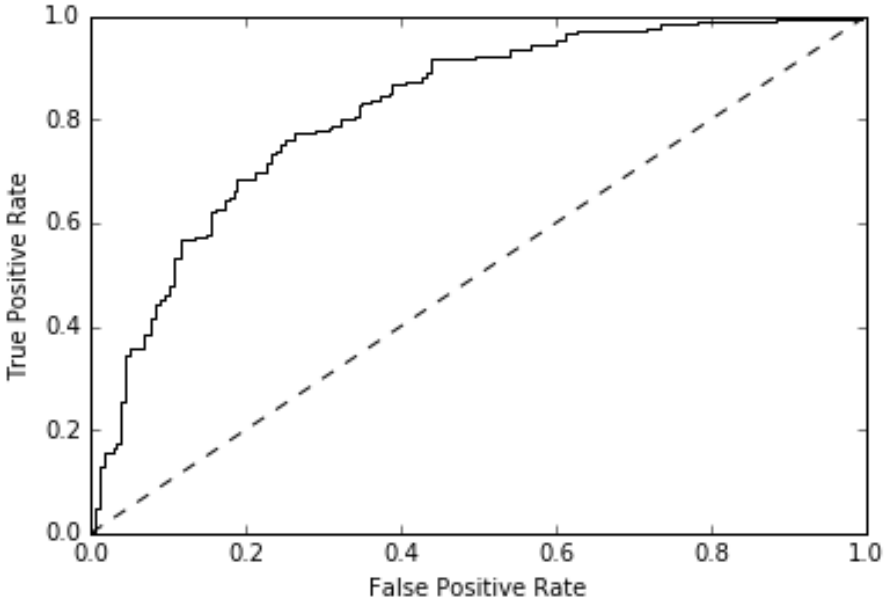
Tablica 25: Usporedba modela neuronske mreže

nosti svih modela veće od 0,8, sva tri promatrana modela smatramo izvrsno prilagođenima. Na osnovu iskustva pretpostavlja se da je skuplje dobiti novog klijenta nego uložiti resurse u zadržavanje već postojećih klijenata. Stoga je cilj minimizirati grešku tipa 1 jer ona predstavlja zanemarivanje neaktivnog klijenta. Greška tipa 2 predstavlja trošenje resursa na zadržavanje



Slika 3: Funkcije troška

klijenta koji će ostati aktivan. Kako iz tablice 25 vidimo da MLP 3 ima najveću AUC vrijednost i najmanju grešku tipa 1, a iz slike 4 ne uočavamo problem s izgledom ROC krivulje modela MLP3, uzimamo njega kao najbolji model neuronske mreže.



Slika 4: ROC krivulja modela MLP3

Sada na istom skupu podataka odljev klijenata modelirajmo logističkom regresijom i uspo-

redimo dobivene rezultate s modelom neuronske mreže.

4.4 Model logističke regresije za procjenu odljeva igrača u online klađenjima

Model 1 kreiran je s 18 varijabli koje su se pokazale značajnima u tablicama 5 i 6. Zbog multikolinearnosti između varijabli odzadnjeggubitka, odzadnjegdobitka i odzadnjeakcije te varijable BrojKladjenjaUZadnjemMjesecu i varijable BrojKladjenjaUZadnjemTjednu kreiramo model 2 s varijablama koje opisuju iznos izgubljen zbog klađenja na izvlačenje loto brojeva (Lost_LOTTO), iznos izgubljen zbog klađenja uživo (Lost_INTERNETLIVE), iznos izgubljen prilikom klađenja na sportske događaje (Lost_SPORT), broj oklada na utrke pasa (BetCount_INTERNETDOGS), broj oklada na klađenje uživo (BetCount_INTERNETLIVE), broj oklada na izvlačenje loto brojeva (BetCount_LOTTO), broj oklada na sportske događaje (BetCount_SPORT), broj fiksnih oklada na klađenje uživo (FixedBetCount_INTERNETLIVE), broj fiksnih oklada na sportske događaje (FixedBetCount_SPORT), broj nefiksnih oklada na izvlačenje loto brojeva (UnfixedBetCount_LOTTO), broj nefiksnih oklada na sportske događaje (UnfixedBetCount_SPORT), odzadnjeggubitka, BrojDanaOdregistracijeDoZadnjeAktivnosti, MaksimalanBrojDanaIzmedjuKladjenja i BrojKladjenjaUZadnjemMjesecu. Korištenjem step procedure iz paketa MASS tražimo model s najmanjom AIC vrijednosti. Taj model za nezavisne varijable koristi iznos izgubljen zbog klađenja uživo (Lost_INTERNETLIVE), broj oklada na utrke pasa (BetCount_INTERNETDOGS), broj oklada na sportske događaje (BetCount_SPORT), odzadnjeggubitka i BrojDanaOdregistracijeDoZadnjeAktivnosti. Nazovimo ga model 3.

	Resid. Df	**Resid. Dev**	**Df**	**Deviance**	**p(>|Chi|)**
Model3	832	837,13			
Model2	825	833,70	7	3,4284	0,8427

Tablica 26: Usporedba modela 2 i modela 3

Anova usporedba modela 3 s modelom 2 dana je u tablici 26. Rezultati pokazuju ne značajnu razliku (p-vrijednost=0,8427). Stoga ćemo dalje promatrati rezultate koje daje model 3.

Procijenjeni parametri modela 3 dani su u tablici 27.

U tablici 28 dani su parametri modela zaokruženi na četiri decimale. b predstavlja logaritam šanse, e^b šansu da klijent bude neaktivan u odnosu na to da bude aktivan, a $\frac{1}{e^b}$ šansu da bude aktivan u odnosu na to da bude neaktivan. Pomoću njih interpretiramo model. Šansa da će klijent biti neaktivan povećava se s povećanjem broja dana od posljednjeg gubitka, a smanjuje se povećanjem iznosa izgubljenog za klađenje uživo, broja oklada na utrke pasa, broja oklada na sportske događaje i broja dana od registracije do zadnje aktivnosti.

Vrijednost procijenjenog parametra uz varijablu Lost_INTERNETLIVE iznosi -0,0019 što znači da ako se iznos izgubljen prilikom klađenja uživo poveća za 1, logaritam šanse da klijent bude neaktivan smanjuje se za 0,0019. Odnosno, jediničnim povećanjem tog iznosa smanjuje se šansa da klijent bude neaktivan u odnosu na to da bude aktivan za 0,9981.

Vrijednost procijenjenog parametra uz varijablu BetCount_INTERNETDOGS iznosi $-0,0037$ što znači da ako se broj oklada na utrke pasa poveća za 1, logaritam šanse da klijent bude ne-

	Parametar	**St. grška**	**z vrijednost**	**p-vrijednost**
(Intercept)	-1,5361135	0,2917380	-5,265	1,4e-07
Lost_INTERNETLIVE	-0,0018693	0,0011145	-1,677	0,0935
BetCount_INTERNETDOGS	-0,0036939	0,0016775	-2,202	0,0277
BetCount_SPORT	-0,0002142	0,0001083	-1,978	0,0479
odzadnjeggubitka2	1,5397020	0,2865126	5,374	7,7e-08
odzadnjeggubitka3	2,3654337	0,2874097	8,230	<2e-16
odzadnjeggubitka4	3,9149944	0,3369503	11,619	<2e-16
BrojDanaOdregistracije-DoZadnjeAktivnosti	-0,0013588	0,0006301	-2,157	0,0310

Tablica 27: Procjena parametara modela 3

aktivan smanjuje se za 0,0037. Odnosno, jediničnim povećanjem broja oklada na utrke pasa smanjuje se šansa da klijent bude neaktivan u odnosu na to da bude aktivan za 0,9963.

Vrijednost procijenjenog parametra uz varijablu BetCount_SPORT iznosi $-0,0002$ što znači da ako se broj oklada na sportske događaje poveća za 1, logaritam šanse da klijent bude neaktivan smanjuje se za 0,0002. Odnosno, jediničnim povećanjem broja oklada na sportske događaje smanjuje se šansa da klijent bude neaktivan u odnosu na to da bude aktivan za 0,9998.

	b	e^b	$\frac{1}{e^b}$
(Intercept)	-1,5361	0,2152	4,6465
Lost_INTERNETLIVE	-0,0019	0,9981	1,0019
BetCount_INTERNETDOGS	-0,0037	0,9963	1,0037
BetCount_SPORT	-0,0002	0,9998	1,0002
odzadnjeggubitka2	1,5397	4,6632	0,2144
odzadnjeggubitka3	2,3654	10,6487	0,0939
odzadnjeggubitka4	3,915	50,1489	0,0199
BrojDanaOdregistracijeDoZadnjeAktivnosti	-0,0014	0,9986	1,0014

Tablica 28: Parametri modela 3

Baznu kategoriju za varijablu odzadnjeggubitka čine klijenti koji su zadnji put izgubili prije manje od sedam dana. Prema tome, šansa da bude neaktivan onaj klijent kojem je od zadnjeg gubitka prošlo između 7 i 36 dana povećava se za 4,6632 u odnosu na onog klijenta koji je zadnji put izgubio prije manje od 7 dana. Šansa da bude neaktivan onaj klijent kojem je od zadnjeg gubitka prošlo između 36 i 78 dana, odnosno za kojeg nemamo podatak o toj varijabli, povećava se za 10,6487 u odnosu na onog klijenta koji je zadnji put izgubio prije manje od 7 dana. Šansa

da bude neaktivan onaj klijent kojem je od zadnjeg gubitka prošlo više od 78 dana povećava se za 50,1489 u odnosu na onog klijenta koji je zadnji put izgubio prije manje od 7 dana.

Vrijednost procijenjenog parametra uz varijablu BrojDanaOdregistracijeDoZadnjeAktivnosti iznosi $-0,0014$ što znači da ako se broj dana od klijentove registracije do zadnje aktivnosti poveća za 1, logaritam šanse da klijent bude neaktivan smanjuje se za 0,0014. Odnosno, jediničnim povećanjem broja dana smanjuje se šansa da klijent bude neaktivan u odnosu na to da bude aktivan za 0,9986.

Matrica konfuzije modela 3

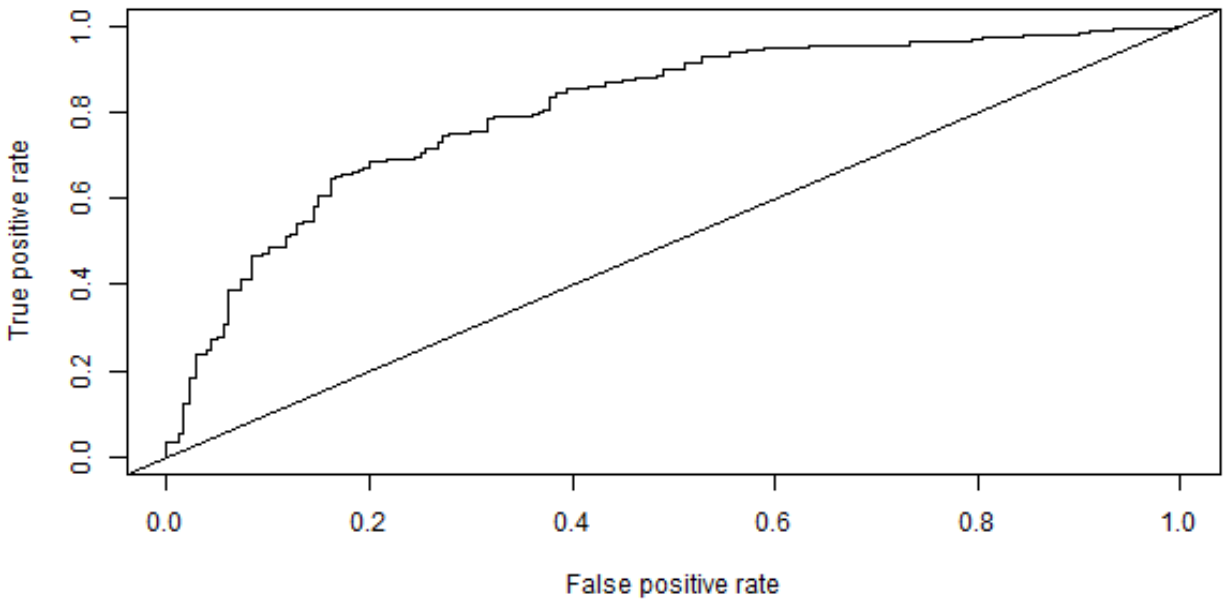
	Predviđeni aktivni	Predviđeni neaktivni
Stvarno aktivni	132	48
Stvarno neaktivni	50	130

Matrica konfuzije modela 3 u postotcima

	Predviđeni aktivni	Predviđeni neaktivni
Stvarno aktivni	37%	13%
Stvarno neaktivni	14%	36%

Tablica 29: Matrica konfuzije modela 3

Model 3 izrađen u ovom poglavlju potrebno je validirati na skupu podataka za testiranje. S obzirom na to da imamo jednak broj aktivnih i neaktivnih klijenata u modelima smo postavili vrijednost cut-off granica na 0,5, pri čemu klijente s vjerojatnošću manjom od 0,5 procjenjujemo kao aktivne, a klijente s vjerojatnošću većom ili jednakom 0,5 procjenjujemo kao neaktivne.



Slika 5: ROC krivulja model 3

AUC vrijednost iznosi 0,805 zbog čega ovaj model svrstavamo u izvrsno prilagođene modele.

4.5 Usporedba modela neuronske mreže i logističke regresije za procjenu vjerojatnosti odljeva igrača

Od varijabli korištenih za kreiranje modela 3 logističke regresije, varijable koje opisuju broj dana od zadnjeg gubitka u kategorijama i broj dana od registracije do zadnje aktivnosti pojavljuju se kao varijable s najvećim težinama u modelu neuronske mreže MLP3. Osim tih varijabli, varijable s najvećim težinama modela neuronske mreže su one koje opisuju prosječan broj dana između dva klađenja, maksimalan broj dana između dva klađenja i one su se pokazale signifikantnima u testovima čiji su rezultati prikazani u tablicama 5 i 6.

Model	**Točnost**	**Greška tipa 1**	**Greška tipa 2**	**AUC-vrijednost**
model neuronske mreže	0,736	0,22	0,31	0,821
model logističke regresije	0,728	0,28	0,27	0,805

Tablica 30: Usporedba modela neuronske mreže i logističke regresije

Usporedbom rezultata modela neuronske mreže i logističke regresije u tablici 30 možemo zaključiti kako bolje rezultate postiže model neuronske mreže jer ima veću točnost i AUC vrijednost, a manju grešku tipa 1. Modeliranje odljeva neuronskom mrežom pokazao se kao dobra alternativa tradicionalnom predikcijskom modelu poput logističke regresije.

5 Zaključak

Logistička regresija traži matematičku vezu između nezavisnih varijabli i zavisne varijable, ali nekada je teško odmah shvatiti prirodu te veze. Neuronske mreže, s druge strane, uočavaju uzorke u vezama. S ciljem predviđanja odljeva klijenata ovaj rad predstavio je nekoliko modela neuronskih mreža i logističke regresije. Za modele neuronskih mreža proveli smo analizu kvalitete na temelju koje se model neuronske mreže MLP3 pokazao kao najprediktivniji. Kreiran je sa 25 čvorova u skrivenom sloju i konstantnom stopom učenja čija je početna vrijednost zadana s 0,2 i momentum 0,9. U modelu MLP3 varijable s najvećim težinama su iznos osvojen za klađenje na sportske događaje, broj oklada na ATG, broj dana od zadnjeg gubitka u kategorijama, prosječan broj dana između dva klađenja i maksimalan broj dana između dva klađenja. Analizom modela logističke regresije najboljim se pokazao model s najmanje nezavisnih varijabli, tj. model 3 te je on detaljnije opisan u radu. Model 3 za nezavisne varijable koristi iznos izgubljen zbog klađenja uživo, broj oklada na utrke pasa, broj oklada na sportske događaje, broj dana od posljednjeg gubitka i broj dana od registracije do posljednje aktivnosti.

Slično kao Coussement i De Bock u [6] u modelu logističke regresije značajnima se pokazuju varijable koje opisuju vrijeme zadnjeg gubitka i gubitak u jednom mjesecu za klađenje uživo. Kao i u prethodnim istraživanjima, model neuronske mreže dao je bolje rezultate od modela logističke regresije. Prednost modela logističke regresije nad modelom neuronske mreže leži u interpretabilnosti parametara. Oba modela imaju potencijal za primjenu. Postoji prostor za nadograđivanje i modifikaciju modela kako bi ga se optimiziralo i proširilo područje djelovanja.

Za daljnje istraživanje zanimljivo bi bilo promotriti kakve bi rezultate ostvarili stablima odlučivanja kao u [8] ili algoritmima učenja ansamblom kao u [6]. Kako u bazi podataka 76% klijenata barem pola ukupnog iznosa plaćenog za klađenje uplaćuje upravo za klađenje na sportske događaje, zanimljivo bi bilo kreirati modele samo na tom podskupu klijenata i vidjeti kako to utječe na izbor varijabli.

Popis slika

1 Struktura neuronske mreže koja ulazni podatak veličine n transformira kroz L slojeva i pridružuje predikcije za pripadanje svakoj od k kategorija 2
2 Primjer ROC krivulje 8
3 Funkcije troška 33
4 ROC krivulja modela MLP3 33
5 ROC krivulja model 3 36

Popis tablica

1 Matrica konfuzije 7
2 Frekvencije varijable ProsloDanaOdZadnjegGubitka s obzirom na varijablu churn 23
3 Frekvencije varijable ProsloDanaOdZadnjegDobitka s obzirom na varijablu churn 24
4 Frekvencije varijable ProsloDanaOdZadnjeAkcije s obzirom na varijablu churn . 24
5 P-vrijednosti χ^2 testa za diskretne slučajne varijable 24
6 P-vrijednosti t-testa za neprekidne slučajne varijable 25
7 Iznos koji je igrač izgubio u jednom mjesecu zbog klađenja uživo s obzirom na aktivnost 26
8 Iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na izvlačenje loto brojeva s obzirom na aktivnost 26
9 Iznos koji je igrač izgubio u jednom mjesecu zbog klađenja na sportske događaje s obzirom na aktivnost 26
10 Broj oklada na utrke pasa u jednom mjesecu s obzirom na aktivnost 27
11 Broj oklada na klađenje uživo u jednom mjesecu s obzirom na aktivnost 27
12 Broj oklada na izvlačenje loto brojeva u jednom mjesecu s obzirom na aktivnost 27
13 Broj oklada na sportske događaje u jednom mjesecu s obzirom na aktivnost . . . 28
14 Broj fiksnih oklada na klađenje uživo u jednom mjesecu s obzirom na aktivnost . 28
15 Broj fiksnih oklada na sportske događaje u jednom mjesecu s obzirom na aktivnost 28
16 Broj nefiksnih oklada na izvlačenje loto brojeva u jednom mjesecu s obzirom na aktivnost 29
17 Broj nefiksnih oklada na sportske događaje u jednom mjesecu s obzirom na aktivnost 29
18 Broj dana od registracije do zadnje aktivnosti s obzirom na aktivnost 29
19 Maksimalan broj dana koji prođe između dva klađenja s obzirom na aktivnost . 30
20 Broj klađenja u zadnjem tjednu s obzirom na aktivnost 30
21 Broj klađenja u zadnjem mjesecu s obzirom na aktivnost 30
22 Matrica konfuzije MLP1 31
23 Matrica konfuzije MLP2 32
24 Matrica konfuzije MLP3 32
25 Usporedba modela neuronske mreže 32
26 Usporedba modela 2 i modela 3 34
27 Procjena parametara modela 3 35
28 Parametri modela 3 35

29 Matrica konfuzije modela 3 36
30 Usporedba modela neuronske mreže i logističke regresije 37

Literatura

[1] L. J. Bain, M. Engelhardt, *Introduction to Probability and Mathematical Statistics*, Pacific Grove, Duxbury/Thomson Learning, 1991.

[2] M. Benšić, *Predavanja za kolegij Statistika*, https://www.mathos.unios.hr/images/homepages/mirta/statistika/sve1.pdf

[3] M. Benšić, N. Šuvak, *Primijenjena statistika*, Sveučilište J.J. Strossmayera, Odjel za matematiku, 2013.

[4] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press Inc., New York, 1995.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[6] K. Coussement, K. W. De Bock, *Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning*, Journal of Business Research, Vol. 66(2013.), 1629-1636

[7] A. J. Dobson, A. G. Barnett, *An introduction to generalized linear models*, CRC Press, Boca Raton, 2018.

[8] B. Huang, M. T. Kechadi, B. Buckley, *Customer churn prediction in telecomunications*, Expert Systems with Applications, Vol. 39(2012.), 1414-1425

[9] M. R. Ismail, M. K. Awang, M. N. A Rahman, M. Makhtar, *A Multi-Layer Perceptron Approach for Customer Churn Prediction*, International Journal of Multimedia and Ubiquitous Engineering, Vol. 10(2015.), 213-222

[10] W. McCulloch, W. Pitts, *A Logical Calculus of Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics, Vol. 5(1943.) 115-133

[11] F. F. Reichheld, Phil Schefter, *E-Loyalty: Your Secret Weapon on the Web*, Harvard Business Review, 78:4 (2000.), 105–113

[12] R. Scitovski, N. Truhar, Z. Tomljanović, *Metode optimizacije*, Sveučilište Josipa Jurja Strossmayera, Odjel za matematiku, 2014.

[13] H. Šimović, A. Bajo, M. Primorac, M. Davidović, F. Jelavić, *Tržište igara na sreću u Hrvatskoj: financijsko poslovanje i fiskalni učinak*, Fiscus, Vol. 9(2019.), 1-32

[14] https://scikit-learn.org/stable/modules/neural_networks_supervised.html

[15] European Gaming and Betting Association, 2017. Most Popular Online Gambling Activities, https://www.egba.eu/resource-post/eu28-most-popular-online-gambling-activities-2017/, preuzeto: 1.12.2019.

[16] European Gaming and Betting Association, 2017. Online Gambling As Share Of Total Gambling Activity, https://www.egba.eu/resource-post/eu28-online-gambling-as-share-of-total-gambling-activity-2017/, preuzeto: 1.12.2019.

[17] Oesterreichische Nationalbank, *Validating rating models, Guidlines on credit risk management*, https://www.oenb.at

Sažetak

Online tržište igara na sreću vrlo je profitabilno. Zbog velike konkurentnosti važno je efektivno zadržati klijente. Svrha predikcije odljeva klijenata je na temelju ponašanja u prošlosti identificirati klijente koji će s velikom vjerojatnošću prestati s klađenjem kod priređivaća igara na sreću. Cilj rada je usporediti modeliranje odljeva neuronskom mrežom s prihvaćenim statističkim modelom logističke regresije. U prvom dijelu rada objašnjene su struktura i učenje neuronske mreže. U drugom dijelu rada predstavljeni su generalizirani linearni modeli i logistička regresija. Obzirom na vrstu modela, međusobno su uspoređeni rezultati tri modela neuronske mreže i tri modela logističke regresije. Nakon toga uspoređen je model neuronske mreže s modelom logističke regresije. Model neuronske mreže ostvario je bolje rezultate od modela logističke regresije.

Ključne riječi: odljev klijenata, online klađenje, neuronska mreža, logistička regresija

Customer churn prediction in the online gambling using neural networks and logistic regression

Summary

The online gambling industry is one of the most profitable branches of the entertainment business. Intense competition makes it very important to effectively retain clients. The purpose of churn prediction is to identify clients with high probability to leave online gambling company based on their past behavior. The aim of this thesis is to critically compare a neural network technique with the established statistical technique of logistic regression for churn prediction. In the first part of the thesis, structure and learning of neural networks were explained. In the second part we introduced generalized linear models and logistic regression. After creating three models using neural networks and three models using logistic regression, the results were compared within techniques. After that, comparison across techniques was made. It was shown that neural network model performed better than logistic regression model.

Key words: churn, online gambling, neural network, logistic regression

Životopis

Rođena sam 6. studenog 1995. godine u Slavonskom Brodu gdje sam završila osnovnu školu "Bogoslav Šulek". Nakon osnovne škole upisujem Ekonomsko-birotehničku školu u Slavonskom Brodu koju završavam 2014. godine. Iste godine obrazovanje nastavljam na Odjelu za matematiku Sveučilišta Josipa Jurja Strossmayera u Osijeku na preddiplomskom studiju matematike kojeg završavam 2017. godine i stječem naziv prvostupnice matematike s temom završnog rada *Simetrične, bisimetrične i persimetrične matrice* pod mentorstvom doc.dr.sc. Darije Marković. U jesen 2017. godine upisujem diplomski studij matematike, smjer Financijska matematika i statistika. Trenutno sam zaposlena kao asistent za aktuarske poslove u službi aktuarskih poslova Uniqa osiguranja u Zagrebu.