# Logistička regresija u predikciji odljeva korisnika mobilne igre

**Šarlija, Anđela**

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ANĐELA ŠARLIJA

# LOGISTIČKA REGRESIJA U PREDIKCIJI ODLJEVA KORISNIKA MOBILNE IGRE

DIPLOMSKI RAD

Split, rujan 2022.

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ODJEL ZA MATEMATIKU

# LOGISTIČKA REGRESIJA U PREDIKCIJI ODLJEVA KORISNIKA MOBILNE IGRE

DIPLOMSKI RAD

Neposredna voditeljica:
dr. sc. Ana Perišić

Mentorica:
doc. dr. sc. Tea Martinić Bilać

Studentica:
Anđela Šarlija

Split, rujan 2022.

# Uvod

Regresijska analiza integralna je komponenta bilo koje analize podataka u kojoj se opisuje veza između jedne (ili više) nezavisnih varijabli i jedne zavisne varijable. Cilj ovakve analize pronalazak je najboljeg modela koji opisuje promatranu vezu. Često je slučaj da je zavisna varijabla dihotomna, odnosno da poprima točno dvije različite vrijednosti. U tom slučaju moguće je primijeniti logističku regresiju koja je u posljednje vrijeme, u različitim područjima, postala standardna metoda analiziranja podataka dihotomnih zavisnih varijabli.

U prvom dijelu diplomskog rada obrađena je teorijska pozadina logističke regresije. Rad započinjemo definiranjem osnovnih pojmova te kratkim uvodom u logističku regresiju koji pretpostavlja postojanje točno jedne nezavisne varijable. Nakon toga, povećanjem broja nezavisnih varijabli, detaljno obrađujemo pojam multivarijatne logističke regresije. Navodimo oblik modela multivarijatne logističke regresije te metodu za procjenu parametara i testiranje značajnosti nezavisnih varijabli modela. Također, obrađena je i interakcija između nezavisnih varijabli te interpretacija procijenjenih parametara. Posljednji dio Poglavlja 2 daje predloženu strategiju za izgradnju logističkog modela koja koristi stepwise proceduru. Nakon procjene modela potrebno je ocijeniti prediktivnu moć modela. U Poglavlju 3 navodimo mjere kojima je navedeno moguće ocijeniti i to koristeći logističku regresiju kao kla-

sifikator.

U drugom dijelu diplomskog rada provodimo istraživanje odljeva korisnika mobilne igre pomoću, prethodno obrađene statističke metode, logističke regresije. Cilj analize jest detekcija igrača za kojeg je izgledno da će kroz neko vrijeme odustati od igre. Određivanje navedenog vremena rezultiralo je s 20 različitih definicija zavisne varijable koju promatramo. Na raspolaganju je dano 19 različitih nezavisnih varijabli te je konačni cilj istraživanja pronaći što bolji multivarijatni logistički model. Istraživanje je podijeljeno na dva dijela. Prvi dio obrađuje univarijatnu, a drugi dio multivarijatnu logističku regresiju. Drugi dio rezultira procijenjenim modelima za četiri odabrane definicije odljeva korisnika.

# Sadržaj

**Uvod** **iii**

**Sadržaj** **v**

**1 Uvod u model** **1**

1.1 Osnovni pojmovi . . . . . 1

1.2 Uvod u logističku regresiju . . . . . 2

**2 Multivarijatna logistička regresija** **4**

2.1 Multivarijatni regresijski model . . . . . 4

2.2 Procjena parametara . . . . . 5

2.3 Testiranje značajnosti varijabli . . . . . 9

2.4 Procjena logita . . . . . 10

2.5 Interpretacija parametara modela . . . . . 11

2.6 Interakcija . . . . . 15

2.7 Izgradnja modela . . . . . 16

**3 Logistička regresija kao klasifikator** **20**

3.1 Procjena prediktivne moći klasifikatora . . . . . 21

3.2 Klasifikacijske tablice kod logističke regresije . . . . . 24

**4 Predikcija odljeva korisnika** **26**

4.1 Opis problema . . . . . 26

4.2 Opis skupa podataka i varijabli . . . . . 27

4.3 Univarijatni logistički modeli . . . . . 32

4.4 Multivarijatni logistički modeli . . . . . 36

**Zaključak** **43**

**Literatura** **44**

# Poglavlje 1

# Uvod u model

## 1.1 Osnovni pojmovi

**Definicija 1.1** *Neka je* $(\Omega, \mathcal{F})$ *izmjeriv prostor i* $\mathcal{P}$ *familija vjerojatnosnih mjera na* $(\Omega, \mathcal{F})$*. Tada je uređena trojka* $(\Omega, \mathcal{F}, \mathcal{P})$ *statistička struktura.*

**Definicija 1.2** *Neka je na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *dan slučajni vektor* $\boldsymbol{Y}\colon \Omega \to \mathbb{R}^n$*. Za fiksni* $\theta \in \Theta$ *označimo s* $F(\cdot;\theta)$ *funkciju distribucije od* $\boldsymbol{Y}$ *u odnosu na vjerojatnost* $P_\theta$*. Familiju* $\mathcal{P}' = \{F(\cdot;\theta) \mid \theta \in \Theta\}$ *nazivamo statističkim modelom, a za vektor* $\boldsymbol{Y}$ *kažemo da pripada tom statističkom modelu.*

**Definicija 1.3** *Neka je* $(\Omega, \mathcal{F}, \mathcal{P})$ *statistička struktura. Slučajan uzorak duljine* $n$ *ili* $n$*-dimenzionalni slučajni uzorak je niz* $X_1, X_2, ..., X_n$ *slučajnih varijabli (vektora) na* $(\Omega, \mathcal{F})$ *takvih da su nezavisne i jednako distribuirane u odnosu na svaku vjerojatnost* $\mathbb{P} \in \mathcal{P}$*.*

**Definicija 1.4** *Neka je* $A \in \mathcal{F}$ *promatrani događaj i* $\pi = \mathbb{P}(A)$*. Tada broj* $\omega = \frac{\pi}{1-\pi}$ *nazivamo izglednost događaja* $A$*.*

## 1.2 Uvod u logističku regresiju

Regresijske metode jedne su od osnovnih metoda bilo koje analize podataka u kojima se opisuje veza između zavisne varijable i jedne (ili više) nezavisnih varijabli. U slučaju kada je zavisna varijabla kategorijalna, i to dihotomna, vrlo često se koristi logistička regresija. Prije svega, bitno je naglasiti cilj ovakve metode: pronalazak što boljeg modela koji opisuje vezu između zavisne varijable (varijable odziva) i nezavisnih varijabli (prediktora).

U ovom uvodnom dijelu pretpostavimo da je zavisna varijabla Y dihotomna (poprima vrijednosti 0 i 1) te da je dana samo jedna nezavisna varijabla X. Tada se radi se o univarijatnoj (ili jednostavnoj) logističkoj regresiji.

Bilo koji regresijski problem, pa tako i promatrani, za cilj ima predvidjeti srednju vrijednost nezavisne varijable za danu vrijednost zavisne varijable, odnosno $\mathbb{E}(Y \mid x)$. U slučaju kada zavisna slučajna varijabla poprima vrijednosti 0 ili 1, onda je $0 \leq \mathbb{E}(Y \mid x) \leq 1$. Jednostavnosti radi označimo $\pi(x) = \mathbb{E}(Y \mid x)$. Univarijatni logistički model koji promatramo sljedećeg je oblika:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \tag{1.1}$$

Transformacija modela $\pi(x)$, koji je središnji dio proučavanja logističke regresije, naziva se logit transformacija i sljedećeg je oblika:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x. \tag{1.2}$$

Primijetimo da je logit transformacija $g(x)$ linearna te može poprimiti vrijednosti između $-\infty$ i $+\infty$.

Cilj je pronaći procjene parametara $\beta_0$ i $\beta_1$ modela koje označavamo s $\hat{\beta}_0$ i $\hat{\beta}_1$ redom. Parametri modela procjenjuju se na temelju uzorka. Metoda koju ćemo koristiti za procjenu parametara logističkog modela (1.2) je metoda

maksimalne vjerodostojnosti. Time ćemo dobiti procjene parametara koji maksimiziraju vjerojatnost dobivanja opaženog uzorka.

Nakon procjene parametara potrebno je ispitati značajnost varijabli, značajnost modela te ako je model značajan od interesa je interpretirati parametre modela. U idućem poglavlju dana je generalizacija modela navedenog u ovom poglavlju.

# Poglavlje 2

# Multivarijatna logistička regresija

U prethodnom poglavlju dan je kratak uvod u logističku regresiju pri čemu je promatran logistički regresijski model koji sadrži jednu nezavisnu varijablu. U ovom poglavlju promatramo slučaj gdje zavisna varijabla poprima vrijednosti 0 ili 1, ali je dano $p \in \mathbb{N} \setminus \{1\}$ nezavisnih varijabli, odnosno prediktora. Ovakav logistički model naziva se multivarijatni logistički regresijski model.

## 2.1 Multivarijatni regresijski model

Pretpostavimo da su slučajne varijable $X_1, X_2, ..., X_p$, $p \in \mathbb{N}\setminus\{1\}$, međusobno nezavisne i neka je $\mathbf{X}' = (X_1, ..., X_p)$ slučajan vektor. Označimo $\mathbb{P}(Y = 1 \mid \mathbf{x}) = \pi(\mathbf{x})$. Logit multivarijatnog logističkog modela dan je izrazom

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{2.1}$$

a multivarijatni logistički model dan je s

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}. \tag{2.2}$$

Ako je nezavisna varijabla diskretna, potrebno je uvesti indikator varijable (*"dummy"* varijable). Pretpostavimo da je nezavisna varijabla $X_j$ diskretna za neki $j = 1, 2, ..., p$. Neka varijabla $X_j$ poprima $k \in \mathbb{N}$ različitih vrijednosti: $L_1, L_2, ..., L_k$. U tom slučaju potrebno je uvesti $k - 1$ indikator varijabli. U ovom slučaju je logit multivarijatnog logističkog modela dan izrazom

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \sum_{l=1}^{k-1} \beta_{jl} D_{jl} + ... + \beta_p x_p$$

pri čemu su $D_{jl}$ indikator varijable konstruirane na sljedeći način:

$$D_{jl} = \begin{cases} 1, & x_j = L_l \\ 0, & \textit{inače} \end{cases},$$

$l = 1, 2, ..., k - 1$.

## 2.2 Procjena parametara

U ovom poglavlju prikazat ćemo postupak procjene parametara u logističkoj regresiji metodom maksimalne vjerodostojnosti koju označavamo MLE (eng. *Maximum likelihood estimation*). Prvo, uvedimo osnovne teorijske pojmove.

**Definicija 2.1** *Statistika na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *je svaka slučajna varijabla (vektor)* $T\colon \Omega \to \mathbb{R}^d$ *takva da postoji* $n \in \mathbb{N}$ *i* $n$*-dimenzionalni slučajni uzorak* $(X_1, ..., X_n)$ *na* $(\Omega, \mathcal{F}, \mathcal{P})$ *te izmjerivo preslikavanje* $t\colon \mathbb{R}^n \to \mathbb{R}^d$ *takvo da je* $T = t(X_1, ..., X_n)$.

### 2.2. Procjena parametara

Neka je $\mathbb{X} = (X_1, ..., X_n)$ slučajan uzorak iz modela $\mathcal{P} = \{F(\cdot; \theta) \mid \theta \in \Theta\}$. Na osnovu zadanog uzorka želimo procijeniti vrijednost parametra $\theta$, ili općenito, neke njegove funkcije $\tau(\theta) \in \tau(\Theta) \subseteq \mathbb{R}^k$.

**Definicija 2.2** *Procjenitelj od $\tau(\theta)$ je statistika $T = t(\mathbb{X}) = t(X_1, ..., X_n)$ u $\mathbb{R}^k$.*

Ako je $\mathbf{x} = (x_1, ..., x_n)$ realizacija slučajnog uzorka $\mathbb{X}$, tada je vjerodostojnost funkcija $l \colon \Theta \to \mathbb{R}$,

$$l(\theta \mid \mathbf{x}) = l(\theta) := f_{\mathbb{X}}(\mathbf{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

**Definicija 2.3** *Statistika $\widehat{\theta} = \widehat{\theta}(\mathbb{X})$ je procjenjitelj maksimalne vjerodostojnosti (MLE) ako vrijedi*

$$l(\widehat{\theta} \mid \mathbb{X}) = \max_{\theta \in \Theta} l(\theta \mid \mathbb{X}).$$

Pretpostavimo da je dan uzorak od $n$, $n \in \mathbb{N}$, nezavisnih opservacija $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$. Pri tome je $y_i$ vrijednost zavisne varijable te $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})$ vrijednosti nezavisnih varijabli za $i - tog$ člana u uzorku, $i = 1, 2, ..., n$. Nadalje pretpostavimo da zavisna varijabla poprima vrijednosti 0 ili 1 što predstavlja odsustvo, odnosno prisustvo nekog obilježja. Označimo s $\boldsymbol{\beta'} = (\beta_0, \beta_1, ..., \beta_p)$ vektor parametara koji pripadaju promatranom modelu. Doprinos funkciji vjerodostojnosti za uređeni par $(\mathbf{x_i}, y_i)$ je $\pi(\mathbf{x_i})$ ako je $y_i = 1$ te $1 - \pi(\mathbf{x_i})$ ako je $y_i = 0$, $i = 1, ..., n$. Dakle, doprinos funkciji vjerodostojnosti za bilo koji uređeni par $(\mathbf{x_i}, y_i)$, $i = 1, 2, ..., n$, možemo zapisati u obliku

$$\pi(\mathbf{x_i})^{y_i}(1 - \pi(\mathbf{x_i}))^{1-y_i}. \tag{2.3}$$

### 2.2. Procjena parametara

Kako su $(\mathbf{x_i}, y_i)$, $i = 1, 2, ..., n$, međusobno nezavisni, funkcija vjerodostojnosti je produkt svih izraza danih u (2.3), odnosno:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x_i})^{y_i}(1 - \pi(\mathbf{x_i}))^{1-y_i}. \tag{2.4}$$

Do procjene parametara $\beta_j$, $j = 0, 1, ..., p$ dolazimo metodom maksimalne vjerodostojnosti. No, jednostavnosti radi, maksimizirat ćemo logaritmiranu funkciju vjerodostojnosti koju nazivamo log-vjerodostojnost:

$$L(\boldsymbol{\beta}) = \ln(l(\boldsymbol{\beta})) = \sum_{i=1}^{n} \big\{ y_i \ln[\pi(\mathbf{x_i})] + (1 - y_i) \ln[1 - \pi(\mathbf{x_i})] \big\}. \tag{2.5}$$

Naime, budući da je logaritmirana funkcija strogo rastuća injekcija, maksimizacija funkcije vjerodostojnosti $l$ ekvivalentna je maksimizaciji funkcije log-vjerodostojnosti. Da bismo pronašli $\boldsymbol{\beta}$ koji maksimizira izraz (2.5), parcijalno deriviramo (2.5) po $\beta_i$, $i = 0, 1, ..., p$, te izjednačimo s nula. Na taj način dobijemo $p + 1$ izraza:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x_i})] = 0, \tag{2.6}$$

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(\mathbf{x_i})] = 0, \tag{2.7}$$

gdje $j = 1, 2, ..., p$. Izrazi (2.6) i (2.7) su nelinearni te do procjena koeficijenata $\beta_i$, $i = 0, 1, ..., p$, dolazimo iterativnim metodama. Procjenu parametra $\boldsymbol{\beta}$ označit ćemo s $\hat{\boldsymbol{\beta}}$, a procjene parametara $\beta_i$, $i = 0, 1, ..., p$, s $\hat{\beta}_i$, $i = 0, 1, ..., p$, redom. Nadalje, procijenjene vjerojatnosti multivarijatnog logističkog modela su $\hat{\pi}(\mathbf{x_i})$, $i = 1, 2, ..., n$, koje dobijemo uvrštavanjem procijenjenih koeficijenta $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ u model dan s (2.2).

Nakon što smo procijenili parametre promatranog modela, procijenit ćemo varijancu i kovarijancu procijenjenih koeficijenata što je usko vezano uz teoriju procjene maksimalne vjerodostojnosti. Procjene su dobivene iz matrice

parcijalnih derivacija drugog reda funkcije log-vjerodostojnosti. Parcijalne derivacije imaju sljedeći oblik:

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i), \tag{2.8}$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1-\pi_i) \tag{2.9}$$

za $j, l = 0, 1, ..., p$ gdje $\pi_i$ označava $\pi(\mathbf{x_i})$. Označimo s $\mathbf{H}(\boldsymbol{\beta})$ matricu dimenzija $(p+1) \times (p+1)$ koja sadrži izraze dane u (2.8) i (2.9). Matricu $\mathbf{I}(\boldsymbol{\beta}) = -\mathbf{H}(\boldsymbol{\beta})$ nazivamo opažena informacijska matrica. Varijance i kovarijance procijenjenih koeficijenata dobivene su upravo iz inverza ove matrice. Označimo $\mathrm{Var}(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$.

Za procjenu varijance i kovarijance koristit ćemo zapise $\widehat{\mathrm{Var}}(\hat{\beta}_j)$ i $\widehat{\mathrm{Cov}}(\hat{\beta}_j, \hat{\beta}_l)$ gdje $j, l = 0, 1, ..., p$. Nadalje, procjenu standardne pogreške procjene označit ćemo s

$$\widehat{\mathrm{SE}}(\hat{\beta}_j) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}$$

za $j = 0, 1, ..., p$. Koristan zapis informacijske matrice modela je $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X'VX}$ gdje su

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{2.10}$$

i

$$\boldsymbol{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}. \tag{2.11}$$

Nakon procjene parametara modela i procjene standardne pogreške procijenjenih parametara možemo procijeniti pouzdane intervale parametara modela. Rubne točke $100(1-\alpha)\%$ pouzdanog intervala za parametar $\beta_j$, $j = 0, 1, ..., p$, su

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \widehat{SE}(\hat{\beta}_j)$$

gdje je $z_{1-\frac{\alpha}{2}}$ $(1-\frac{\alpha}{2})$-kvantil standardne normalne distribucije.

## 2.3 Testiranje značajnosti varijabli

Nakon procjene parametara možemo testirati hipotezu o značajnosti varijabli našeg modela. Jedan od pristupa povezan je sa sljedećim pitanjem: "Govori li nam više model s uključenom varijablom ili model bez promatrane varijable o zavisnoj varijabli?" Odgovor na ovo pitanje moguće je dobiti usporedbom opaženih i predviđenih vrijednosti dobivenih iz modela s promatranom varijablom i bez promatrane varijable. Za usporedbu koristit ćemo sljedeći statistiku

$$D = -2 \ln \left[ \frac{\textit{vjerodostojnost procijenjenog modela}}{\textit{vjerodostojnost saturiranog modela}} \right], \tag{2.12}$$

gdje je saturirani model model sa savršenim *fitom*, tj. onaj koji sadrži onoliko parametara koliko je elemenata u promatranom uzorku. Statistika D naziva se statistika odstupanja (devijanca). Vrijedi

$$D = -2 \sum_{i=1}^{n} \left[ y_i \ln(\frac{\hat{\pi}_i}{y_i}) + (1-y_i) \ln(\frac{1-\hat{\pi}_i}{1-y_i}) \right],$$

gdje je $\hat{\pi}_i = \hat{\pi}(\mathbf{x_i})$.

Statistika $G$ koja bilježi promjenu u devijanci D ovisno o uključivanju nezavisne varijable u model dana je izrazom:

$$G = D(\textit{model bez varijable}) - D(\textit{model s varijablom}). \tag{2.13}$$

Kako je vjerodostojnost saturiranog modela jednaka za obje devijance u izrazu (2.13), G možemo izraziti na sljedeći način:

$$G = -2\ln\left[\frac{\textit{vjerodostojnost modela bez varijable}}{\textit{vjerodostojnost modela s varijablom}}\right]. \tag{2.14}$$

Na temelju statistike G moguće je testirati značajnost varijabli. Uz pretpostavku istinitosti hipoteze $H_0 : \beta_j = 0$ statistika G ima asimptotsku $\chi^2$ distribuciju uz $p$ stupnjeva slobode.

Testiranje značajnosti varijabli moguće je provesti temeljem Waldovog testa. Testna statistika dana je s

$$W_j = \frac{\hat{\beta}_j}{\widehat{\mathrm{SE}}(\hat{\beta}_j)},$$

gdje je $j = 1, 2, ..., p$. Uz pretpostavku istinitosti nul-hipoteze $H_0 : \beta_j = 0$, ove statistike su asimptotski normalno distribuirane za sve $j = 1, 2, ..., p$.

## 2.4 Procjena logita

Procjenu logita $\hat{g}(\boldsymbol{x})$ dobijemo jednostavno tako što u logit (2.1) uvrstimo vrijednosti procijenjenih parametra $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$. Dakle, procjena logita dana je s

$$\hat{g}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p.$$

Moguće je procjenu izvršiti pouzdanim intervalom, a za to nam je potrebna varijanca procjenjitelja logita koja je dana s

$$\widehat{\mathrm{Var}}[\hat{g}(\mathbf{x})] = \sum_{j=0}^{p} x_j^2 \widehat{\mathrm{Var}}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k \widehat{\mathrm{Cov}}(\hat{\beta}_j, \hat{\beta}_k).$$

Alternativan zapis za procjenu logita je

$$\hat{g}(\mathbf{x}) = \mathbf{x}'\hat{\boldsymbol{\beta}}$$

gdje je vektor $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ vektor procijenjitelja koeficijenata $\beta_0, \beta_1, ..., \beta_p$, a $\mathbf{x}' = (x_0, x_1, ..., x_p)$ pri čemu je $x_0 = 1$. Sada možemo varijancu procjenjitelja logita prikazati jednostavnije i to pomoću matričnog zapisa na sljedeći način:

$$\widehat{\text{Var}}[\hat{g}(\mathbf{x})] = \mathbf{x}'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}$$

gdje su matrice $\mathbf{X}$ i $\mathbf{V}$ definirane kao u (2.10) i (2.11). Nakon procjene logita možemo odrediti $100(1-\alpha)\%$ pouzdan interval za logit, a njegove granice su

$$\hat{g}(\mathbf{x}) \pm z_{1-\frac{\alpha}{2}} \widehat{SE}[\hat{g}(\mathbf{x})],$$

gdje je $\widehat{SE}[\hat{g}(\mathbf{x})] = \sqrt{\widehat{\text{Var}}[\hat{g}(\mathbf{x})]}$, a $z_{1-\frac{\alpha}{2}}$ $(1-\frac{\alpha}{2})$-kvantil standardne normalne distribucije.

## 2.5 Interpretacija parametara modela

Nakon procjene modela naglasak s računanja i procjene parametara prelazi na interpretaciju procijenjenih parametara. Pretpostavimo da je procijenjen multivarijatni logistički model te da su nezavisne varijable modela značajne, bilo u statističkom ili kliničkom smislu. Interpretirati bilo koji trenirani model zahtijeva odgovor na sljedeće pitanje: "Što procijenjeni koeficijenti u modelu govore o pitanju koje je motiviralo samu studiju?" Procijenjeni koeficijenti nezavisnih varijabli predstavljaju promjenu u zavisnoj varijabli pri jediničnom povećanju u istoj toj nezavisnoj varijabli. Dakle, interpretacija uključuje dva problema: određivanje veze između zavisne i nezavisne varijable te prikladno definiranje jedinične promjene u zavisnoj varijabli.

Kod logističke regresije promjena u zavisnoj varijabli određena je promjenom u logitu, odnosno promatramo razliku $g(\mathbf{x}+1)-g(\mathbf{x})$. Ispravna interpretacija značila bi da možemo dati značenje razlici između ova dva logita. Da

bismo to detaljno objasnili, promotrimo različite slučajeve u ovisnosti o vrsti nezavisne varijable čiji koeficijent interpretiramo. Jednostavnosti radi, promotrimo slučaj univarijatnog logističkog modela, a onda ćemo razmatranja proširiti na slučaj multivatijatnog logističkog modela. Dakle, i dalje pretpostavljamo da zavisna varijabla $Y$ poprima vrijednosti 0 ili 1 te pretpostavimo da imamo je dana točno jedna nezavisna varijabla $X$.

Neka je nezavisna varijabla $X$ dihotomna i neka poprima vrijednosti 0 ili 1. Razlika u logitu za subjekta kojem je $x = 1$ i subjekta za kojeg je $x = 0$ je $g(1) - g(0) = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$. Dakle, u ovom slučaju razlika u logitu jednaka je $\beta_1$. Da bismo interpretirali ovaj rezultat, moramo se upoznati s mjerom koja se naziva omjer izglednosti. Slučajevi koji se mogu dogoditi prikazani su u tablici 2.1.

| Nezavisna varijabla Y | x=1 | x=0 |
|---|---|---|
| y=1 | $\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$ | $\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ |
| y=0 | $1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$ | $1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$ |

Tablica 2.1: Vrijednosti logističkog modela kad je nezavisna varijabla dihotomna.

Omjer izglednosti, u oznaci OR, definiran je kao omjer izglednosti za subjekte kojima je $x = 1$ i $x = 0$ redom, odnosno

$$\text{OR} = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]}. \tag{2.15}$$

Ako zamijenimo izraze za logistički model u (2.15) onako kako je navedeno

u tablici (2.1), onda dobijemo sljedeće:

$$\begin{aligned}\text{OR} &= \frac{(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}})/(\frac{1}{1+e^{\beta_0+\beta_1}})}{(\frac{e^{\beta_0}}{1+e^{\beta_0}})/(\frac{1}{1+e^{\beta_0}})} = \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = \\ &= e^{(\beta_0+\beta_1)-\beta_0} = \\ &= e^{\beta_1}.\end{aligned}$$

Dakle, ako je nezavisna varijabla dihotomna i poprima vrijednosti 0 ili 1, onda je veza između omjera izglednosti i regresijskog koeficijenta

$$\text{OR} = e^{\beta_1}.$$

Upravo ova jednostavna veza između omjera izglednosti i regresijskog koeficijenta jedan je od razloga zašto je logistička regresija tako moćan alat u analitičkim istraživanjima. Omjer izglednosti široko se primjenjuje i predstavlja aproksimaciju odnosa izglednosti nastupanja događaja od interesa u grupi subjekta za koje je $x = 1$.

**Primjer 2.4** *Neka y predstavlja prisustvo ili odsustvo raka pluća te neka je x dihotomna varijabla koja poprima vrijednost 1 ukoliko je osoba pušač, a 0 ukoliko nije. Pretpostavimo da je* $\widehat{OR} = 2$. *Takav rezultat ukazuje da je izglednost raka pluća 2 puta veća u populaciji pušača nego u populaciji nepušača.*

**Primjer 2.5** *Neka y predstavlja prisustvo ili odsustvo srčane bolesti i ako x predstavlja redovnu tjelovježbu (*$x = 1$ *ukoliko osoba redovno vježba), onda* $\widehat{OR} = 0.5$ *ukazuje da je izglednost prisustva srčane bolesti* 0.5 *puta manja u populaciji onih koji provode redovnu tjelovježbu.*

Pretpostavimo sada da nezavisna varijabla $X$ poprima točno $k$ vrijednosti, $k \in \mathbb{N}\setminus\{1,2\}$. Pretpostavimo da su $D_1, D_2, ..., D_{k-1}$ indikator varijable

i referentnu grupu označimo s $D$. U ovom slučaju koeficijente intepretiramo tako da uspoređujemo referentnu grupu posebno sa svakom od preostalih grupa. Dakle, koeficijent $\beta_j$ odnosi se na usporedbu referentne grupe i $j$-te grupe, $j = 1, 2, ..., k-1$. Interpretira se analogno kao u dihotomnom slučaju. Ako želimo usporediti dvije kategorije od kojih niti jedna nije referentna, onda gledamo razliku koeficijenata. Točnije, tada je $OR = e^{\beta_j - \beta_l}$ gdje su $\beta_j$ i $\beta_l$ koeficijenti dvaju različitih dizajn varijabli, $j = 1, 2, ..., k-1$. Interpretacija je analogna kao u dihotomnom slučaju osim što referentnu grupu mijenja grupa čija je pripadna indikator varijabla $D_j$.

Pretpostavimo sada da je nezavisna varijabla $X$ kontinuirana. Uz pretpostavku linearnosti logita, vrijedi $g(x+1) - g(x) = \beta_1$, za proizvoljan $x$. Dakle, $\beta_1$ predstavlja promjenu u log-izglednosti pri povećanju vrijednosti varijable $x$ za jednu jedinicu. U praksi često nije od interesa promatrati za jednu jedinicu varijable $x$, već za više jedinica. Zato ćemo obratiti pozornost na promjenu u $c \in \mathbb{N} \setminus \{1\}$ jedinica. Log-izglednost za promjenu u $c$ jedinica dobivena je iz razlike logita $g(x+c) - g(x) = c\beta_1$ i pripadni omjer izglednosti je $OR = e^{c\beta_1}$. Ovo implicira da je pri povećanju od $c$ jedinica varijable $X$, rizik od nastupanja događaja ($y = 1$) veća (ili manja) $e^{c\beta_1}$ puta.

Konačno možemo interpretirati parametre multivarijatnog logističkog modela. Pretpostavimo da je multivarijatni logistički model oblika (2.2), odnosno da imamo $p \in \mathbb{N} \setminus \{1\}$ prediktora. Koeficijent $\beta_j$, $j = 1, 2, ..., p$, interpretiramo analogno kao u univarijatnom slučaju uz naglasak da su svi ostali prediktori fiksni.

## 2.6 Interakcija

Interakcija između nezavisnih varijabli multivarijatnog logističkog modela može biti prisutna u različitim formama. Započnimo sa slučajem kad interakcija ne postoji. Proučimo model koji ima jednu dihotomnu i jednu kontinuiranu varijablu. Ako je veza između kontinuirane varijable i zavisne varijable jednaka za obje kategorije dihotomne varijable, onda ne postoji interakcija između nezavisnih varijabli. Odsustvo interakcije možemo uočiti grafički ako usporedimo grafove logita za dva modela: u jednom fiksiramo jednu kategoriju dihotomne varijable, a u drugom fiksiramo drugu kategoriju. Odsustvo interakcije možemo pretpostaviti ako logiti imaju "gotovo isti" koeficijent smjera. S druge strane, prisustvo interakcije grafički možemo uočiti ako se spomenuti logiti vidno razlikuju. Dakle, kad imamo multivarijatni logistički model bitno je provjeriti postoji li dokaz o postojanju interakcije između nezavisnih varijabli. Ako postoji interakcija između varijabli, onda ne možemo interpretirati koeficijente na već objašnjen način.

Da bismo objasnili interpretaciju u slučaju postojanja interakcije, pretpostavimo da imamo dvije nezavisne varijable, dihotomnu varijablu $F$ i kontinuiranu varijablu $X$. Uz prisustvo interakcije logit promatranog multivarijatnog logističkog modela je oblika

$$g(f, x) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 f \times x.$$

Pretpostavimo da želimo usporediti omjer izglednosti dviju kategorija varijable $F$, $F = f_1$ i $F = f_0$ na razini $X = x$. Tada je

$$g(f_1, x) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x$$

i

$$g(f_0, x) = \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x.$$

Nadalje, izračunajmo log-izglednost:

$$\begin{aligned} ln[\text{OR}(\text{F}=f_1,\ \text{F}=f_0,\ \text{X}=\text{x})] &= g(f_1,x) - g(f_0,x) = \\ &= (\beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x) - \\ &- (\beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x) = \\ &= \beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0). \end{aligned} \tag{2.16}$$

Potom izračunamo omjer izglednosti $OR = e^{\beta_1(f_1-f_0)+\beta_3 x(f_1-f_0)}$. Primijetimo da se izraz za log-izglednost dan u (2.16) ne odnosi na jedan koeficijent. Umjesto toga, uključuje dva koeficijenta, razliku u vrijednostima varijable $F$ i interakcijsku varijablu. Procjena omjera izglednosti dobivena je zamjenom koeficijenata modela u (2.16) s odgovarajućim procjenama parametara.

## 2.7 Izgradnja modela

U prethodnim potpoglavljima fokusirali smo se na procjenu multivarijatnog logističkog modela za kojeg smo imali točno određene nezavisne varijable. U praksi je često da na raspolaganju postoji više nezavisnih varijabli koje potencijalno mogu biti uključene u model. Stoga, moramo razviti strategiju za izgradnju što "boljeg" modela. Da bismo to postigli potrebno je odrediti strategiju odabira nezavisnih varijabli te razviti metode za određivanje adekvatnosti modela. Uspješno modeliranje kompleksnog skupa podataka nema unaprijed određenu šablonu, već je spoj statističkih metoda, iskustva i teorijske podloge. Kriterij za uključivanje varijabli u model varira ovisno o problemu i znanstvenim disciplinama. Tradicionalni pristup statističke izgradnje modela uključuje pronalazak što jednostavnijeg modela koji dobro opisuje podatke. Navodimo niz koraka koje možemo pratiti pri odabiru varijabli koje uključujemo u model. Naravno, moguće je pratiti i drugačije strategije.

1. Proces selekcije započinje univarijatnom analizom svake varijable. Primjerice, za kategorijalne nezavisne varijable koje poprimaju samo nekoliko vrijednosti možemo pripremiti $2 \times k$ kontigencijsku tablicu gdje se 2 odnosi na dvije vrijednosti koje može poprimiti zavisna varijabla (0 ili 1), a $k$ predstavlja broj različitih vrijednosti koje može poprimiti nezavisna kategorijalna varijabla koju proučavamo. Test koji možemo primijeniti je $\chi^2$ test. Nadalje, moguće je procijeniti univarijatni logistički model i ispitati značajnost varijable. Analogno, za kontinuirane varijable moguća univarijatna analiza jest procjena univarijatnog logističkog modela. Tako možemo provesti test značajnosti nezavisne varijable koji smo prethodno obradili. Alternativna analiza jest provođenje t-testa. Univarijatna analiza nezavisnih varijabli korisna je prilikom odluke o uključivanju varijabli u multivarijatni logistički model.

2. Nakon univarijatne analize biramo varijable za multivarijatnu analizu. Općenito, varijable koje su značajno povezane s promatranom zavisnom varijablom uključuju se u multivarijatni model. Na primjer, kao kandidate za multivarijatni model moguće je odabrati varijable koje su značajne u univarijatnim modelima. Moguće je odabrati i manje strog pristup. Primjerice, u multivarijatni model uključiti varijable čiji univarijatni testovi imaju p-vrijednost manju od neke unaprijed određene veličine. Ponekad se u modele uključuju i varijable za koje u univarijatnoj analizi nije dokazana značajna povezanost sa zavisnom varijablom, ali su teorijski povezane sa zavisnom varijablom. Nakon identifikacije varijabli koje uključujemo u model, započinjemo s modelom koji sadrži sve odabrane varijable. Problem koji može nastati s bilo kojim univarijatnim pristupom je da zanemaruje mogućnost da varijable koje su pojedinačno bile slabo povezane sa zavisnom varijablom

mogu biti važni prediktori u kombinaciji s nekim drugim nezavisnim varijablama. Dakle, trebali bismo odabrati razinu značajnosti koja je dovoljno velika da dopusti takvim varijablama uključenje u multivarijatni logistički model. Popularan pristup za selekciju varijabli jest stepwise metoda. Stepwise metoda temeljena je na statističkom algoritmu koji provjerava značajnost varijabli i prema tome se odlučuje hoće li varijabla biti uključena ili isključena iz modela. Razlikujemo dvije osnovne stepwise metode: metoda unaprijed i metoda unatrag. Metoda unaprijed započinje s nul-modelom i dodajemo varijable koje zadovoljavaju unaprijed određen statistički kriterij. Metoda unatrag započinje sa uključenim svim prediktorima te postupno izbacujemo varijable koje ne zadovoljavaju kriterij. Moguće je provesti i kombinaciju ovih dviju metoda. Korake stepwise metode najlakše je objasniti na primjeru što je obrađeno u posljednjem poglavlju.

3. Nakon odabira varijabli, možemo procijeniti multivarijatni logistički model. Potrebno je ispitati značajnost svake varijable uključene u model. To uključuje univarijatne (Waldove) testove za svaku varijablu modela te usporedbu svakog procijenjenog koeficijenta s odgovarajućim procijenjenim koeficijentom u univarijatnom modelu koji sadrži samo tu određenu varijablu. Nadalje, možemo usporediti multivarijatni model u kojem smo isključili promatranu varijablu i multivarijatni model s uključenom varijablom. Također možemo usporediti procijenjene koeficijente tih dvaju modela. Kod usporedbe procijenjenih koeficijenata problem nastaje ako se oni razlikuju u velikoj mjeri. To nas navodi na zaključak da su jedna ili više isključenih varijabli bile značajne. Ovakav proces traje sve dok ne pronađemo sve varijable koje su značajne. Nakon što smo na ovaj način dobili finalni model, poželjno je postupno

vraćati prethodno isključene varijable u model. Na ovaj način možemo identificirati varijablu koja zasebno nije bila značajno povezana sa zavisnom varijablom, ali ima značajan doprinos zajedno s ovim, već odabranim, varijablama. Krajnji model trećeg koraka često se naziva *"preliminary main effects model"*.

4. Sada kada smo dobili model koji sadrži značajne varijable, trebamo pažjivije proučiti varijable tog modela. Za nezavisne varijable koje poprimaju samo nekoliko vrijednosti ostajemo na univarijatnoj razini, dok za kontinuirane varijable trebamo provjeriti pretpostavku linearnosti logita. U slučaju nelinearnosti logita moguće je provesti različite transformacije nezavisne varijable ili uključiti članove višeg reda. Krajnji model četvrtog koraka često se naziva *"main effects model"*.

5. Nakon profinjenja modela, potrebno je proučiti interakcije između odabranih varijabli. U bilo kojem modelu interakcija između dvije varijable implicira da utjecaj jedne varijable na zavisnu varijablu nije konstantan obzirom na drugu varijablu. Konačna odluka hoćemo li uključiti interakciju u model ovisi o rezultatima analize, ali i o teorijskoj podlozi. Krajnji model petog koraka često se naziva *"preliminary final model"*.

Prije korištenja bilo kojeg procijenjenog modela moramo provjeriti njegovu značajnost. Dakle, ako je konačni model značajan, onda ga možemo koristiti za predviđanje vrijednosti zavisne varijable na temelju onih varijabli koje su uključene u konačni model.

# Poglavlje 3

# Logistička regresija kao klasifikator

Klasifikacija je forma analize podataka koja se bavi izgradnjom modela koji detaljnije opisuje klase ili kategorije unutar promatranih podataka. Spomenuti modeli, koje zovemo klasifikatori, predviđaju kojoj klasi pripada neki podatak.

Klasifikacija se sastoji od dva koraka. Prvi korak je korak učenja u kojem se gradi klasifikacijski model, dok je drugi korak klasifikacijski korak u kojem se izgrađeni model koristi za predviđanje klase promatranih podataka. Pri tome skup podataka podijeljen je na skup za treniranje na kojem se model uči, odnosno gradi, te na skup za testiranje. Skup za treniranje, a time i skup za testiranje, odabran je na slučajan način. Kako prilikom treniranja modela imamo uvid u to o kojoj se klasi radi, govorimo o nadziranom učenju. Prvi korak procesa klasifikacije započinje određivanjem funkcije koja predviđa klasu za dani podatak. U drugom koraku provjeravamo točnost klasifikacije, odnosno procjenjujemo prediktivnu moć klasifikatora. O tome ćemo detaljnije u sljedećem potpoglavlju.

## 3.1 Procjena prediktivne moći klasifikatora

Procjenu prediktivne moći klasifikatora radimo na skupu za testiranje koji je nezavisan od skupa za treniranje na kojem se model učio. Točnost klasifikatora ispitujemo usporedbom stvarne i predviđene klase svakog elementa iz skupa za testiranje. Ako se točnost klasifikatora smatra prihvatljivom, onda klasifikator možemo koristiti za klasifikaciju nekih novih podataka za koje nemamo informaciju o klasi, već je klasifikator taj koji ju određuje. Točnost klasifikatora nije jedina mjera za određivanje je li neki klasifikator "dobar" pa ćemo se upoznati s još nekim mjerama prediktivne moći, a to su osjetljivost i specifičnost. Prije svega uvodimo potrebnu nam terminologiju. Podatke dijelimo na pozitivne i negativne pri čemu su pozitivni oni čija je klasa ona klasa od interesa. Pretpostavimo da koristimo klasifikator na skupu za testiranje kojem su stvarne klase poznate. Označimo s P broj pozitivnih podataka, a N broj negativnih podataka. Za svaki podatak uspoređujemo predviđenu klasu sa stvarnom klasom. Navedimo još četiri dodatna pojma.

- Stvarno pozitivni (u oznaci TP): Podaci koji su ispravno prepoznati kao pozitivni od strane klasifikatora.
- Stvarno negativni (u oznaci TN): Podaci koji su ispravno prepoznati kao negativni od strane klasifikatora.
- Lažno pozitivni (u oznaci FP): Podaci koji su negativni, ali su prepoznati kao pozitivni od stane klasifikatora.
- Lažno negativni (u oznaci FN): Podaci koji su pozitivni, ali su prepoznati kao negativni od stane klasifikatora.

Navedeni pojmovi sažeti su u tablici (3.1) za slučaj kad postoje točno dvije klase.

| | **Predviđena klasa** | | |
|---|---|---|---|
| **Stvarna klasa** | da | ne | Ukupno |
| da | TP | FN | P |
| ne | FP | TN | N |
| Ukupno | P' | N' | P+N |

Tablica 3.1: Konfuzijska matrica.

Konfuzijska matrica (matrica zabune, *"confusion matrix"*) koristan je alat za analiziranje uspješnosti klasifikatora. TP i TN odnose se na ispravno prepoznate klase podataka, dok FP i FN odnose se na neispravno prepoznate klase podataka. Ako imamo $m \in \mathbb{N} \setminus \{1\}$ klasa, onda je konfuzijska matrica reda $m$. Ako klase označimo s $C_j$, $j = 1, 2, ..., m$, onda element na mjestu $(i, k)$, $i, k = 1, 2, ..., m$, označava broj elementata koji su klase $C_i$, a klasifikator ih je prepoznao kao klasu $C_k$. Da bi klasifikator imao visoku prediktivnu sposobnost, većina podataka izvan dijagonale bi trebalo biti blizu nula.

Sada možemo definirati osnovne mjere za vrednovanje modela koristeći uvedene oznake. Definiramo točnost (*"accuracy"*) kao omjer ispravno klasificiranih podataka i ukupnog broja podataka

$$\textit{točnost} = \frac{\text{TP} + \text{TN}}{\text{P+N}}.$$

Ponekad se točnost naziva i stopa prepoznavanja (*"recognition rate"*) te se točnost može koristiti kada je distribucija klasa relativno balansirana. Korisna mjera povezana s točnosti je stopa greške (*"misclassification rate"*), koju jednostavno računamo kao *1 − točnost*. Vrijedi

$$\textit{stopa greške} = \frac{\text{FP} + \text{FN}}{\text{P} + \text{N}}.$$

### 3.1. Procjena prediktivne moći klasifikatora

Ako koristimo skup za treniranje, umjesto skupa za testiranje, da bismo procijenili stopu greške promatranog klasifikatora, onda je ova mjera poznata kao *resubstitution error*.

Pretpostavimo da klase nisu jednoliko distribuirane te da je prisutan problem disbalansa klasa u smislu da je klasa od interesa rijetka. Točnije, distribucija podataka odražava značajnu većinu negativne klase nasuprot pozitivne klase. Ono što se može dogoditi jest to da je točnost velika, ali da nije dobra mjera valjanosti testa. To jest, može se dogoditi da klasifikator točno klasificira negativne podatke, ali netočno klasificira pozitivne podatke što nam svakako nije cilj. Umjesto mjere točnosti potrebne su druge mjere koje govore koliko dobro klasifikator prepoznaje pozitivne podatke te s druge strane koliko dobro klasifikator prepoznaje negativne podatke. Mjere koje koristimo u ove svrhe su osjetljivost i specifičnost redom.

Osjetljivost je proporcija pozitivnih podataka koji su ispravno identificirani, dok je specifičnost proporcija negativnih podataka koji su ispravno identificirani. Ove dvije mjere definirane su na sljedeći način:

$$osjetljivost = \frac{TP}{P}$$

i

$$specifičnost = \frac{TN}{N}.$$

Može se pokazati da vrijedi sljedeće:

$$točnost = osjetljivost * \frac{P}{P+N} + specifičnost * \frac{N}{P+N}.$$

## 3.2 Klasifikacijske tablice kod logističke regresije

Pretpostavimo da je dan multivarijatni logistički model čija zavisna varijabla, kao i prethodno, poprima vrijednosti 0 i 1. Intuitivan način za sažimanje rezultata ovakvog logističkog modela je pomoću klasifikacijskih tablica. Da bismo konstruirali pripadnu klasifikacijsku tablicu, moramo odrediti graničnu vrijednost $c \in [0, 1]$ koja određuje pripada li neki podatak klasi 0 ili 1. Dakle, uspoređivat ćemo procijenjene vjerojatnosti logističkog modela s određenom graničnom vrijednošću $c$. Ako je procijenjena vjerojatnost veća od $c$, onda podatak klasificiramo u klasu 1, u suprotnom u klasu 2. Najčešće korištena granična vrijednost $c$ jednaka je 0.5. U ovakvom pristupu, procijenjene vjerojatnosti korištene su za predviđanje kojoj grupi pripada neki podatak.

Osjetljivost i specifičnost oslanjanju se na jedinstvenu graničnu vrijednost $c$ koja služi za klasifikaciju je li podatak pozitivan ili negativan. Za opis točnosti klasifikacije koristi se ROC krivulja (eng. *Receiver operating characteristic curve*), odnosno površina ispod ROC krivulje, u oznaci AUC (eng. *Area under the curve*). ROC krivulja, originalno dolazeći iz teorije detekcije signala, daje vjerojatnost da primatelj detektira istinit signal (osjetljivost) kad je pristutan lažni signal (1-specifičnost). Površina ispod ROC krivulje, koja poprima vrijednosti između 0 i 1, je mjera sposobnosti modela da razlikuje subjekte za koje je $Y = 1$ od onih kojima je $Y = 0$.

Pretpostavimo da je dan izgrađen logistički model i da nas zanima predviđanje zavisne varijable za dane vrijednosti nezavisnih varijabli. Odredimo graničnu vrijednost $c$. Ako tražimo optimalni $c$ za naš model, onda biramo $c \in [0, 1]$ takav da maksimizira osjetljivost i specifičnost. Ovakav $c$ dobijemo tako što definiramo dvije funkcije $f_1$ i $f_2$: argument funkcije

$f_1$ je osjetljivost, a argument funkcije $f_2$ specifičnost. Konačno, traženi $c$ je onaj za koji je $f_1(c) = f_2(c)$. Dakle, ROC krivulja je ona koja na $x$ osi sadrži 1-specifičnost, a na $y$ osi osjetljivost. Točnije, svaku točku krivulje dobijemo tako što za svaku graničnu vrijednost $c \in [0, 1]$ računamo $X_c = (1 - specifičnost, osjeljivost)$. Često korišteno pravilo za ocjenu uspješnosti klasifikatora na temelju AUC-a je:

- Ako je $AUC = 0.5$, onda govorimo o bezvrijednoj klasifikaciji (isto kao da bacamo simetričan novčić).
- Ako je $0.7 \leq AUC < 0.8$, onda govorimo o prihvatljivoj klasifikaciji.
- Ako je $0.8 \leq AUC < 0.9$, onda govorimo o odličnoj klasifikaciji.
- Ako je $AUC \geq 0.9$, onda govorimo o izvanrednoj klasifikaciji.

# Poglavlje 4

# Predikcija odljeva korisnika

## 4.1 Opis problema

Odljev (eng. *"churn"*) korisnika analizira se u različitim industrijama, najčešće u telekomunikacijskom sektoru, ali također i u području bankarstva, osiguranja, pružatelja internetskih usluga, uslužnih djelatnosti, novinarskog sektora, online igara itd. Među mnogim tehnikama za modeliranje predviđanja odljeva korisnika najpopularnije su stabla odlučivanja, neuronske mreže i logistička regresija. Široke mogućnosti statističkih metoda, metoda rudarenja podataka te strojnog učenja za predviđanje odljeva korisnika s jedne strane zadaje teškoće pri odabiru modela. Usprkos činjenici da je modeliranje odljeva korisnika opsežno istraženo, ne postoji generalni konsenzus o izvedbi tehnika modeliranja istog. Budući da usporedbe uspješnosti različitih tehnika često daju različite rezultate, pitanje koju tehniku koristiti jest otvoreno istraživačko pitanje. Tehnika korištena u ovom diplomskom radu jest logistička regresija.

Problem odljeva korisnika kojim ćemo se mi baviti vezan je uz industriju mobilnih igara pa se u tom slučaju problem svodi na detekciju igrača

za koje je izgledno da će u neko, unaprijed određeno vrijeme, odustati od igre. Odabir spomenutog vremena poprilično je subjektivan i često je temeljen na heurističkim pravilima određenima od strane istraživača. Nadalje, trajanje vremena predviđanja razlikuje se ovisno o problemu. Obično se za predviđanje budućeg ponašanja koriste podaci prikupljeni u prethodnih 14 ili 30 dana.

Obrada podataka i izgradnja logističkih modela obrađena je u statističkom programu R-u.

## 4.2 Opis skupa podataka i varijabli

Podaci koje koristimo za izgradnju logističkog modela prikupljeni su u suradnji s hrvatskim IT poduzećem Nanobitom. Skup podataka konstruiran je s ciljem simuliranja problema u kojem bi se na specifičan datum evaluirao rizik odljeva korisnika na temelju interakcije korisnika s igrom. Aktivni igrači koji su pod visokim rizikom odustajanja od igre mogu potencijalno ostati u igri ako ih na neki način pokušamo zadržati. U našem slučaju aktivni igrači su birani na slučajan način, bez obzira na to koliko dugo igraju igru. Skup podataka formiran je slučajnim odabirom 10 000 korisnika koji su imali zabilježen događaj na određeni datum. U svrhu eliminacije igrača koji nisu imali određene prošle aktivnosti i događaje, provedeno je filtriranje podataka. Kriterij za zadržavanje igrača u uzorku jest da je igrač imao barem jednu sesiju u igri, da je prešao barem jedan level i da ima ukupno barem 5 minuta trajanja sesije. Promatranje ne samo zabilježenih događaja, već i spomenutog kriterija o prelasku barem jednog levela opravdava se činjenicom da zabilježen događaj ne mora značiti da je igrač uistinu igrao igru. Na primjer, igrač može pokrenuti igru i ostaviti je otvorenom dok zapravo radi nešto drugo. Proces

filtriranja eliminirao je približno 16% podataka pa je tako preostalo 8415 igrača s oko tri milijuna zabilježenih događaja.

Inicijalni korak rješavanja problema odljeva korisnika je identificiranje "najboljih podataka". Kvaliteta podataka koji se koriste za izgradnju modela bitno utječe na snagu i točnost modela. Dodatno, različite kombinacije varijabli rezultiraju različitom prediktivnom moći modela. U području igara informacije o ponašanju igrača pronalazimo u zapisu radnji igrača, dnevnim prijavama u igru, kupnjama unutar igre, duljini igranja, broju prijeđenih levela itd. Stoga je odabir varijabli na temelju kojih će biti izgrađen model jedan od najvažnijih koraka u procesu izradnje prediktivnih modela. Takozvane RFM (*"Recency-Frequency-Monetary"*) varijable pokazale su se kao dobar izvor ponašanja korisnika općenito. Pri tome se R (*"Recency"*) odnosi na recentnost događaja ("kada je zadnji put zabilježen neki događaj od interesa"). F (*"Frequency"*) odnosi se na frekvenciju ili učestalost događaja (učestalost igranja ili akcija u nekom periodu), dok se M (*"Monetary"*) odnosi na monetarne varijable ili varijable koje bilježe novčane vrijednosti (novac koji je igrač potrošio u igri tijekom nekog perioda). Dodatno, u ovome radu korišten je prošireni RFM model, u oznaci RFM-LIR model. Uz RFM varijable uključene su i varijable vezane uz životni vijek igrača (L=*"lifetime"*), intenzitet igre (I=*intensity*) i akcije provedene u svrhu dobivanja nagrade (*"reward"*).

Varijable su konstruirane koristeći šest osnovnih događaja koji su unaprijed definirani od strane developera: "početak igre", "kraj igre", "prelazak levela", "potrošen novac", "kupnja završena" i "pogledan reklamni video". Kako bismo agregirali više dnevnih podataka u jedan podatak, korišteno je eksponencijalno izglađivanje (EMA, eng. *"Exponential Moving Average"*) s ciljem davanja veće važnosti prilikom računanja prosjeka novijim podacima.

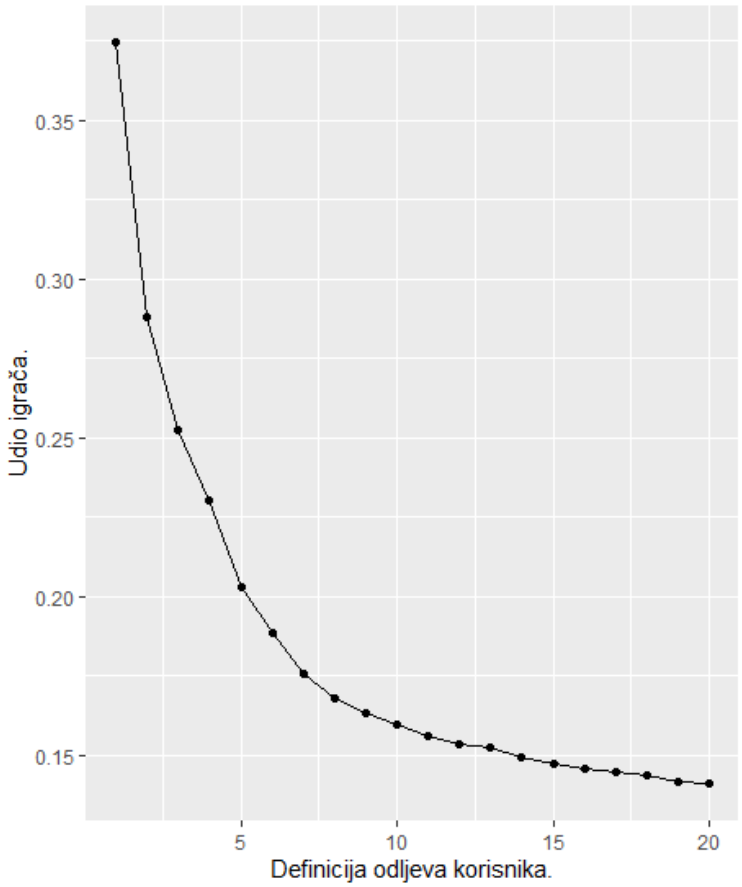
### 4.2. Opis skupa podataka i varijabli

Potencijalne varijable za izradu logističkog modela za predviđanje osipa korisnika dane su u tablici 4.1.

| Kategorija | Varijabla [oznaka] | Transformacija | Učestalost bilježenja |
|---|---|---|---|
| (R) Recency | Broj dana od zadnje sesije [Rec_Ses]<br>Broj dana od zadnjeg prijeđenog levela [Rec_Lvl] | Diskretizirano (imao/nije imao sesiju/ prijeđen level jučer) | Na kraju opservacijskog perioda |
| | Br. dana od zadnje potrošnje valute1 [Rec_C1] | - | Na kraju opservacijskog perioda |
| | Br. dana od zadnje potrošnje valute2 [Rec_C2] | Diskretizirano (imao/nije imao zabilježenu potrošnju valute2 jučer) | Na kraju opservacijskog perioda |
| (F) Frequency | Broj prijeđenih levela [Lvl_Nr]<br>Broj sesija [Ses_Nr]<br>Duljina sesije [Ses_Len] | EMA | Dnevno |
| | Ukupna duljina sesija [LifeSes_Len]<br>Ukupan broj sesija [LifeSes_Nr]<br>Ukupan br. prijeđenih levela [LifeLvl_Nr] | - | Na kraju opservacijskog perioda |
| (M) Monetary | Potrošeno valute1 [C1_Spnt]<br>Potrošeno valute2 [C2_Spnt] | EMA | Dnevno |
| | Životna vrijednost igrača [LTV] | Diskretizirano (potrošač/nije potrošač) | Na kraju opservacijskog perioda |
| (L) Lifetime | Životni vijek u danima [Life] | - | Na kraju opservacijskog perioda |
| (I) Intensity | Prosječna dnevna duljina sesije<br>Prosječni dnevni br. sesija<br>Prosječni dnevni br. prijeđenih levela | - | Na kraju opservacijskog perioda |
| (R) Rewards | Pogledana reklama-monetarna nagrada [Rew_cur] | Diskretizirano gledano/nije gledao | Dnevno |
| | #Pogledana reklama-vremenska nagrada [Rew_sec] | suma | Dnevno |

Tablica 4.1: Popis varijabli zajedno s njihovim svojstvima.

### 4.2. Opis skupa podataka i varijabli




Slika 4.1: Udio churnera.

Prije same izgradnje logističkih modela, potrebno je odrediti definiciju odljeva korisnika. Kod besplatnih igara korisnici ne prekidaju korištenje igre tako što raskidaju neki ugovor, već jednostavno prestanu igrati igru. Dakle, u području mobilnih igara definicija odljeva korisnika obično se oslanja na odustnost igrača u unaprijed određenom vremenskom razdoblju. U slučajnom uzorku igrača koji koristimo u ovoj analizi definiran je predikcijski period duljine 20 dana te je za svakog igrača za svaki dan unutar predikcijskog perioda zabilježeno je li igrač bio prisutan u igri. Dakle, zapravo promatramo 20 različitih definicija u smislu duljine perioda neaktivnosti. Pri tome koris-

timo oznaku $\text{Def}_k$ za definiciju kojom je igrač klasificiran kao churner ako $k$ uzastopnih dana nije imao aktivnosti u igri, $k = 1, 2, ..., 20$. Duljina perioda neaktivnosti $k$ naziva se *churn window*. Na slici 4.1 dan je grafički prikaz udjela churnera obzirom na definiciju. Očekivano se udio churnera smanjuje s povećanjem duljine perioda neaktivnosti. Primijetimo da s porastom promatranog perioda neaktivnosti $k$ opada broj churnera te je neuravnoteženost između klasa churner i ne-churner sve veća.

Konačno, za definiciju odljeva korisnika, od potencijalnih 20 različitih definicija, posebnu pažnju obratit ćemo na: $\text{Def}_1$, $\text{Def}_7$, $\text{Def}_{14}$ i $\text{Def}_{20}$.

Prije same izgradnje modela za predviđanje odljeva korisnika, skup podataka podijeljen je na skup za treniranje i skup za testiranje i to tako da je na slučajan način odabrano 75% podataka za skup za treniranje, a preostali podaci čine skup za testiranje.

Strategija izgradnje logističkih modela oslanja se na identificiranje najjednostavnijeg modela. Prvo ćemo provesti univarijatnu analizu za svaku od četiri odabrane definicije, a potom multivarijatnu analizu koristeći stepwise metodu. Stepwise metoda primijenjena je u statističkom programu R u kojem je kriterij za odabir najboljeg modela Akaikeov informacijski kriterij (AIC). Prilagodba modela ispitana je primjenom Hosmer-Lemeshow testa, mjerama dobivenim iz klasifikacijskih tablica (točnost, osjetjivost i specifičnost) te izračunavanjem površine ispod ROC krivulje (AUC). Poseban naglasak stavljen je na mjerenje osjetljivosti i AUC-a zbog činjenice da točnost nije najbolja mjera u slučaju neuravnoteženih skupova podataka. Zbog problema neuravnoteženosti podataka, provedeno je nasumično poduzorkovanje koje eliminira primjere većinske klase i kao rezultat daje uravnotežene podatke.

## 4.3 Univarijatni logistički modeli

U ovom odjeljku prikazani su rezultati provedene univarijatne logističke regresije. Pri tome je analizirano 19 različitih prediktora (iz 6 kategorija varijabli prikazanih u 4.1) i različite definicije odljeva u smislu različite duljine perioda neaktivnosti (*"churn window size"*). Na slici 4.2 uspoređena je prediktivna snaga univarijatnih modela mjerena površinom ispod ROC krivulje (AUC). Za svaku od 19 nezavisnih varijabli proučavamo kako se kreće AUC na skupu za testiranje pri promjeni definicije odljeva korisnika odnosno pri promjeni duljine perioda neaktivnosti. Pri tome u obzir uzimamo sve definicije ($\text{Def}_1$, $\text{Def}_2$, ..., $\text{Def}_{20}$).

Prediktori vezani za učestalost igre imaju najveću prediktivnu snagu. Pri tome, kratkoročni pokazatelji učestalosti bilježe znatno višu razinu prediktivne snage za kraće duljine perioda dopuštene neaktivnosti. Prema razini prediktivne moći slijede prediktori vezani uz intenzitet i recentnost igre.

Na sličan način analizirana je promjena regresijskog koeficijenta uz nezavisnu varijablu $\beta_1$ svakog od promatranih modela. Na grafovima je koeficijent $\beta_1$ označen s *beta*. To je prikazano na 4.3.

Kao što smo već naveli, koncentirat ćemo se na definicije odljeva korisnika za $\text{Def}_1$, $\text{Def}_7$, $\text{Def}_{14}$ i $\text{Def}_{20}$. U tablici 4.2 navedeni su svi procijenjeni parametri univarijatnih modela za navedene definicije. Da bismo dali konkretan primjer univarijatnog logističkog modela, detaljnije ćemo obraditi dva univarijatna modela: jedan s nezavisnom varijablom koja je kontinuirana i jedan s nezavisnom varijablom koja je kategorijalna. Za definiciju odljeva korisnika uzet ćemo duljinu perioda neaktivnosti od jednog dana.

Proučimo univarijatni logistički model kojem je nezavisna varijabla Lvl_Nr. U ovom slučaju nezavisna varijabla je kontinuirana. Procijenjeni model

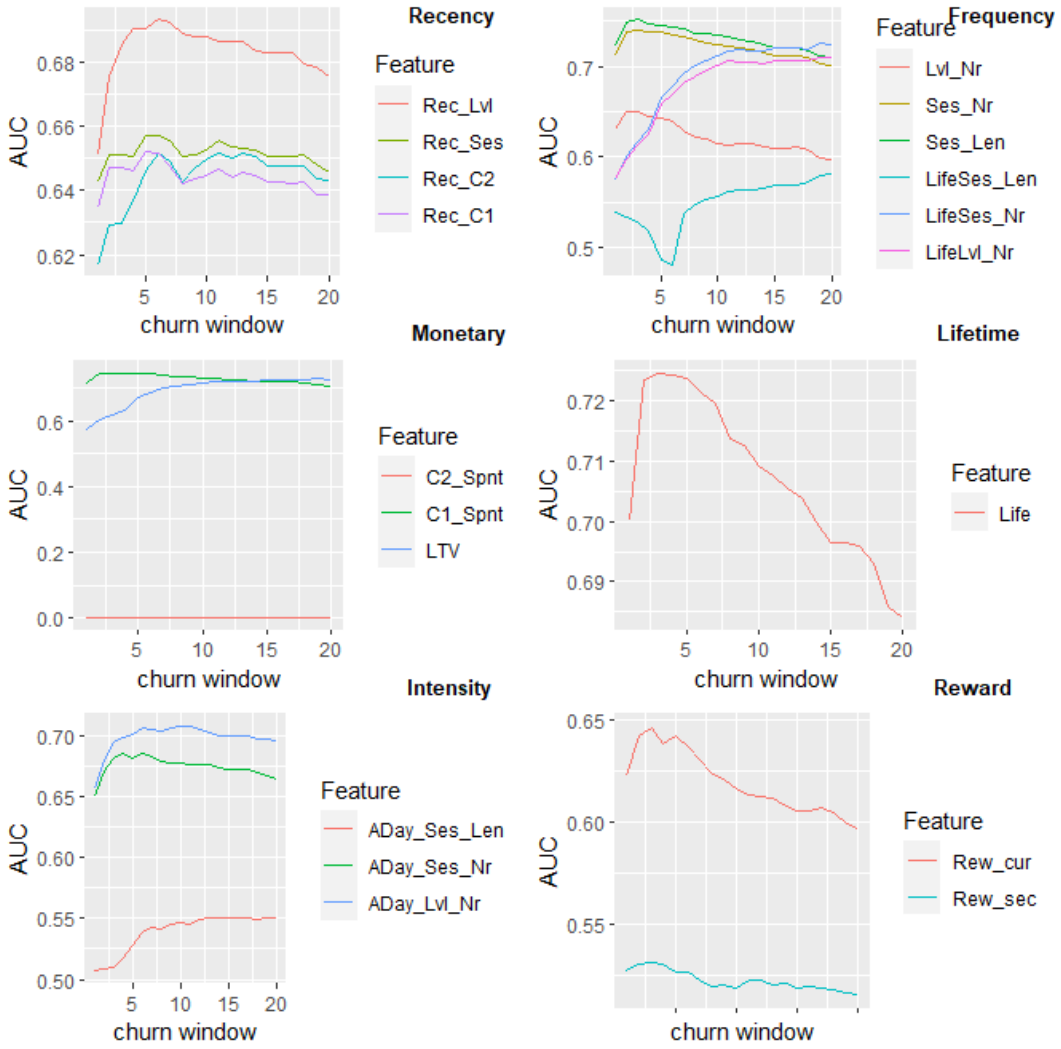
oblika je:

$$g_1(x) = \ln\left[\frac{\pi_1(x)}{1-\pi_1(x)}\right] = 0.1742 + (-0.1870)x.$$

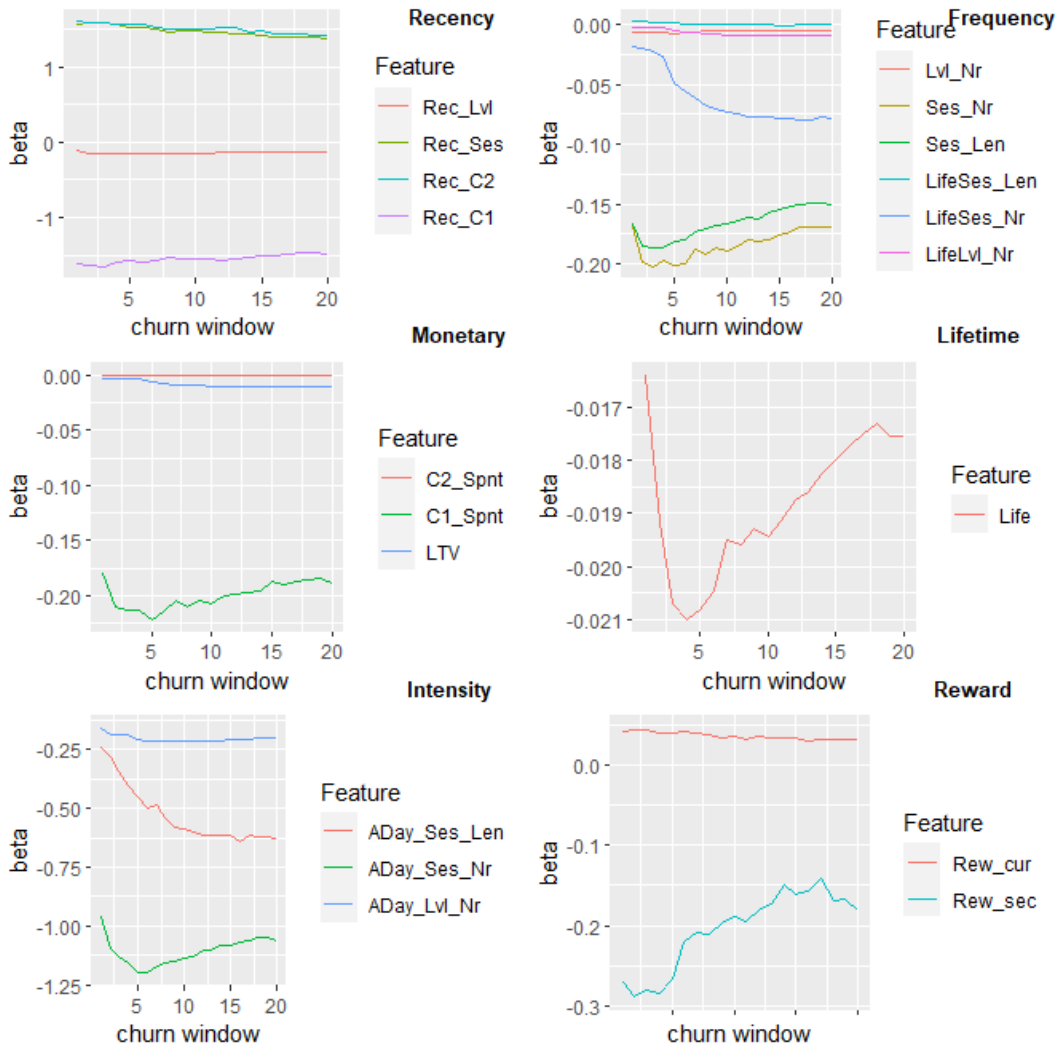
Dakle, $\beta_0 = 0.1742$ i $\beta_1 = -0.1870$. Iz modela lako dobijemo da je

$$\pi_1(x) = \frac{e^{0.1742+(-0.1870)x}}{1+e^{0.1742+(-0.1870)x}}.$$

Testirajmo značajnost nezavisne varijable. Provodimo Waldov test. Vrijed-




Slika 4.2: AUC univarijatnih modela pri povećanju duljine perioda dopuštene neaktivnosti (grupirano prema kategorijama prediktora).




Slika 4.3: Promjena koeficijenta $\beta_1$ pri povećanju prozora (posebno prikazano za svaku varijablu).

nost statistike W jednaka je -19.439, a pripadna p-vrijednost iznosi $2e^{-16}$. Dakle, p-vrijednost je manja od razine značajnosti 1% pa odbacujemo nul-hipotezu i zaključujemo da je nezavisna varijabla značajna. Budući se radi o univarijatnom slučaju, isti test možemo provesti i za značajnost samog modela. Zaključak je da je model značajan. Nadalje, 95% pouzdani interval

za koeficijent $\beta_1$ je $< -0.20588, -0.16817 >$. AUC modela je 0.71209 pa bismo po kriteriju modela kao klasifikatora promatrani model ocijenili kao prihvatljivu klasifikaciju jer je $0.7 \leq AUC < 0.8$. Kako je model značajan, ima smisla interpretirati procijenjeni koeficijent. Kako je $\beta_1 = -0.1870$, to je $\text{OR} = e^{\beta_1} = e^{-0.1870} = 0.8294437$. Dakle, za svaki porast u nezavisnoj varijabli od 1 jedinice izglednost da će igrač biti churner pada 0.8294437 puta. Isto tako možemo interpretirati procijenjeni pouzdani interval. Tako dobivamo da za za svaki porast u nezavisnoj varijabli od jedne jedinice pad izglednosti da će igrač biti churner kreće se između 0.81393 i 0.84521 s 95% pouzdanosti.

| | $\textbf{Def}_1$ | | $\textbf{Def}_7$ | | $\textbf{Def}_{14}$ | | $\textbf{Def}_{20}$ | |
|---|---|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_0$ | $\beta_1$ | $\beta_0$ | $\beta_1$ | $\beta_0$ | $\beta_1$ |
| Rec_Lvl | -0,91337 | 1,57802 | -2,03877 | 1,50929 | -2,207 | 1,42867 | -2,26439 | 1,39284 |
| Rec_Ses | -0,81747 | 1,59128 | -1,93484 | 1,52367 | -2,12012 | 1,47408 | -2,17835 | 1,43212 |
| Rec_C2 | 0,72108 | -1,60887 | -0,45682 | -1,57073 | -0,69208 | -1,51835 | -0,78659 | -1,48239 |
| Rec_C1 | -0,70032 | 0,03945 | -1,70288 | 0,02685 | -1,86974 | 0,02274 | -1,92847 | 0,02102 |
| Lvl_Nr | 0,17418 | -0,18702 | -0,80275 | -0,23145 | -1,04079 | -0,21512 | -1,13399 | -0,20643 |
| Ses_Nr | 0,23992 | -0,17919 | -0,77832 | -0,20816 | -1,02437 | -0,19128 | -1,12007 | -0,18288 |
| Ses_Len | 0,09902 | -0,01787 | -0,84824 | -0,0234 | -1,08893 | -0,0215 | -1,17944 | -0,02067 |
| Life_Ses_Len | -0,31146 | -0,02024 | -0,99712 | -0,07121 | -1,083 | -0,09093 | -1,14294 | -0,09314 |
| Life_Ses_Nr | -0,31904 | -0,00252 | -1,06234 | -0,00828 | -1,16274 | -0,01062 | -1,22648 | -0,01082 |
| lifetime_ch_number | -0,31732 | -0,00274 | -1,01266 | -0,00973 | -1,10023 | -0,0125 | -1,15459 | -0,01297 |
| C2_Spnt | 0,17154 | -0,1983 | -0,78721 | -0,25288 | -1,02257 | -0,23661 | -1,11816 | -0,22631 |
| C1_Spnt | -0,43748 | -0,00754 | -1,46987 | -0,00812 | -1,6684 | -0,00676 | -1,74262 | -0,00603 |
| LTV | -0,44848 | -0,24367 | -1,43272 | -0,50789 | -1,60346 | -0,61236 | -1,66904 | -0,63208 |
| Life | -0,57277 | 0,00217 | -1,54728 | 0,00027 | -1,71166 | $-5e^{-4}$ | -1,77671 | -0,00061 |
| ADay_Ses_Len | 0,08673 | -1,0095 | -0,83318 | -1,34505 | -1,07111 | -1,24658 | -1,15596 | -1,21396 |
| ADay_Ses_Nr | 0,10941 | -0,16756 | -0,77596 | -0,24211 | -0,96426 | -0,2459 | -1,05357 | -0,23858 |
| ADay_Lvl_Nr | 0,01687 | -0,12724 | -0,89041 | -0,17759 | -1,11184 | -0,16859 | -1,19928 | -0,16298 |
| Rew_cur | -0,45025 | -0,26654 | -1,49892 | -0,19921 | -1,69563 | -0,15309 | -1,76159 | -0,16566 |
| Rew_sec | -0,2763 | -0,00105 | -1,29535 | -0,00135 | -1,49901 | -0,00125 | -1,57147 | -0,00123 |

Tablica 4.2: Koeficijenti svih univarijatnih modela za četiri promatrane definicije odljeva korisnika.

Promotrimo sada slučaj univarijatnog logističkog modela kad je nezavisna varijabla kategorijalna. Za primjer možemo uzeti Rec_Lvl (i dalje za definiciju odljeva korisnika promatramo dan 1). U ovom slučaju procijenjeni model je oblika:

$$g_2(x) = \ln\left[\frac{\pi_2(x)}{1-\pi_2(x)}\right] = -0.91337 + 1.57802x. \tag{4.1}$$

Pri tome je

$$\pi_2(x) = \frac{e^{-0.91337+1.57802x}}{1+e^{-0.91337+1.57802x}}.$$

Dakle, procijenjeni koeficijent $\beta_0 = -0.91337$, dok je $\beta 1 = 1.57802$. Kao i u prethodnom primjeru provodimo test značajnosti nezavisne varijable s istim hipotezama. Kao rezultat dobivamo p-vrijednosti koje su "blizu" 0 pa zaključujemo da je varijabla statistički značajna. U ovom slučaju je 95% pouzdani interval za koeficijent $\beta_1$ jednak $< 1.45474, 1.70129 >$. Nadalje, $AUC = 0.64601$ pa bismo ovaj model kao klasifikator ocijenili kao prihvatljiv. Na samom kraju možemo interpretirati procijenjeni koeficijent i interval pouzdanosti. Imamo OR $= e^{\beta_1} = e^{1.57802} = 4.845353$ pa izglednost da je igrač churner 4.845353 puta veća u skupini igrača koji dan prije uzorkovanja nisu imali zabilježenu aktivnost (skupina 1) u odnosu na skupinu igrača koji su imali zabilježenu aktivnost (skupina 0). Interval pouzdanosti nam govori da s pouzdanošću od 95% procjenjujemo kako je izglednost da že igrač biti churner između 4.28337 i 5.481013 puta veća u skupini 1 nego u skupini 0.

## 4.4 Multivarijatni logistički modeli

Promatranje univarijatnih logističkih modela služilo nam je za provođenje multivarijatne logističke analize koja je konačni cilj proučavanja odljeva korisnika. Dakle, cilj nam je pronaći "najbolji" multivarijatni logistički model. Prilikom izgradnje multivarijatnih logističkih modela koristit ćemo sljedeće

nezavisne varijable: C1_Spnt ($i = 1$), Lvl_Nr ($i = 2$), Life ($i = 3$), lifetime_ch_number ($i = 4$), LTV (kategorijalna varijabla, $i = 5$), ADay_Ses_Nr ($i = 6$), Rec_Ses (kategorijalna varijabla, $i = 7$), Rec_C1 ($i = 8$), Rew_cur (kategorijalna varijabla, $i = 9$) i Rew_sec ($i = 10$). Na samom početku multivarijatne analize iz skupa podataka za treniranje isključeni su svi podaci koji odstupaju od medijana za više od 5 interkvartilnih jedinica. Nakon isključivanja izdvojenica skup za treniranje obuhvaća 5526 igrača. Kao metodu za procjenu najboljih multivarijatnih logističkih modela koristimo stepwise metodu i to onu koja koristi kombinaciju metode unaprijed i metode unatrag.

| | Koeficijent | Procjena koef. | Stand. pogreška | z-vrijednost | p-vrijednost |
|---|---|---|---|---|---|
| zavisna var. | $\beta_0$ | 0.39496 | 0.09004 | 4.38667 | $e^{-5}$ |
| C1Spnt | $\beta_1$ | 0.02089 | 0.00569 | 3.67287 | 0.00024 |
| LvlNr | $\beta_2$ | -0.12005 | 0.01823 | -6.58475 | $\approx 0$ |
| Life | $\beta_3$ | -0.00397 | 0.00194 | -2.04724 | 0.04063 |
| lifetime_ch_number | $\beta_4$ | $-6e^{-5}$ | 0.00071 | -0.08516 | 0.93213 |
| LTV | $\beta_5$ | -0.18686 | 0.09592 | -1.94807 | 0.05141 |
| ADay_Ses_Nr | $\beta_6$ | -0.05196 | 0.0158 | -3.28878 | 0.00101 |
| Rec_Ses | $\beta_7$ | 1.01356 | 0.10017 | 10.11836 | $\approx 0$ |
| Rec_C1 | $\beta_8$ | 0.01822 | 0.00846 | 2.15203 | 0.03139 |
| Rew_cur | $\beta_9$ | 0.07706 | 0.08598 | 0.89624 | 0.37012 |
| Rew_sec | $\beta_{10}$ | -0.00028 | 0.00013 | -2.16827 | 0.03014 |

Tablica 4.3: *Full-model* za $\text{Def}_1$.

Prvi model koji procjenjujemo je multivarijatni logistički model u kojem je definicija odljeva korisnika $\text{Def}_1$. Napomenimo da $\beta_i$ označava koeficijent uz varijablu koja je prethodno numerirana s $i$, $i = 1, ..., 10$. Prilikom procjene modela koristili smo metodu s ciljem balansiranja kategorija churnera i nechurnera. Da bismo proveli stepwise proceduru i dobili najbolji model, koji ćemo nazivati finalni model, krenimo od modela koji sadrži sve varijable

koje mogu biti uključene u finalni model. U našem slučaju, radi se o 10 nabrojanih varijabli, a model koji sadrži svih 10 varijabli nazivat ćemo *full-model*. Dobiveni procijenjeni koeficijenti, procjena standardne pogreške, z vrijednosti (za pojedinačne Waldove testove) i pripadne p-vrijednosti za *full-model* dani su u tablici 4.3.

Nadalje, devijanca *nul-modela* je 5659 s 4081 stupnjeva slobode, dok je devijanca procijenjenog modela 5099 s 4071 stupnjeva slobode. Pripadni AIC iznosi 5121. Značajnost pojedinih varijabli promatranog *full-modela* možemo odrediti promatrajući zadnji stupac tablice 4.3. Varijable koje možemo ocijeniti da nisu značajne su one s p-vrijednošću većom od 0.05. Dakle, to su varijable lifetime_ch_number, LTV i Rew_cur. No, finalni model procijenit ćemo pomoću stepwise metode i dalje koristeći statistički program R. Kao kriterij za uključivanje, odnosno isključivanje varijable korišten je Akaikeov informacijski kriterij (AIC). Slijedi prikaz procedure. Na samom početku naveden je AIC *full-modela* te su popisane nezavisne varijable. U svakom koraku stavljen je minus kod varijable koja se isključuje iz modela ili plus kod varijable koja se uključuje u model. Uz to pratimo AIC koji je naveden unutar zagrada pokraj svake promatrane varijable. Ako dobijemo AIC koji je bolji (bolji AIC je onaj koji je manji), onda tu varijablu uključujemo ili isključujemo ovisno o tome je li znak bio plus ili minus, redom. Navedimo dobivene rezultate za svaki korak:

1. korak - svih 10 varijabli je uključeno u model. AIC=5120.86.
    - - lifetime_ch_number (5118.9)
    - - Rew_cur (5119.7)
    - ništa (5120.9)
    - - LTV (5122.7)

- - Life (5123.0)
- - Rec_C1 (5123.6)
- - Rew_sec (5123.6)
- - ADay_Ses_Nr (5129.9)
- - C1Spnt (5132.5)
- - LvlNr (5166.7)
- - Rec_Ses (5227.5)

- AIC=5118.87. Trenutni model sadrži varijable: C1Spnt, LvlNr, Life, LTV, ADay_Ses_Nr, Rec_Ses, Rec_C1, Rew_cur i Rew_sec.

2. korak

- - Rew_cur (5117.7)
- ništa (5118.9)
- + lifetime_ch_number (5120.9)
- - LTV (5120.9)
- - Rec_C1 (5121.6)
- - Rew_sec (5121.6)
- - Life (5125.2)
- - C1Spnt (5130.5)
- - ADay_Ses_Nr (5131.5)
- - LvlNr (5164.7)
- - Rec_ses (5225.6)

- AIC=5117.7. Trenutni model sadrži varijable: C1Spnt, LvlNr, Life, LTV, ADay_Ses_Nr, Rec_Ses, Rec_C1 i Rew_sec.

3. korak

   - ništa (5117.7)
   - + Rew_cur (5118.9)
   - + "lifetime_ch_number" (5119.7)
   - - Rew_sec (5120.1)
   - - LTV (5120.2)
   - - Rec_C1 (5120.3)
   - - Life (5124.4)
   - - ADay_Ses_Nr (5129.9)
   - - C1Spnt (5130.5)
   - - LvlNr (5164.2)
   - - Rec_Ses (5224.2)

4. Finalni model sadrži varijable C1Spnt, LvlNr, Life, LTV, ADay_Ses_Nr, Rec_Ses, Rec_C1 i Rew_sec.

Nakon procjene modela stepwise metodom cilj je pronaći model s uključenim interakcijama. Metoda kojom procjenjujemo model s interakcijama ponovno je stepwise metoda. Za finalni model s interakcijama promotrimo tablicu danu u 4.4. Iz navedene tablice računamo mjere točnosti procijenjenog finalnog modela na skupu za testiranje: točnost koja iznosi 0.66316, specifičnost 0.69868 i osjetljivost 0.62763.

Za preostale definicije odljeva korisnika $\text{Def}_7$, $\text{Def}_{14}$ i $\text{Def}_{20}$ provodimo analognu proceduru, no bez detaljnog navođenja svih koraka. Dakle, na samom početku provodimo proces balansiranja churnera i ne-churnera, zatim procjenjujemo *full-model* s uključenih svih 10 nezavisnih varijabli na kojem

| | 1 | 0 |
|---|---|---|
| 1 | 1281 | 615 |
| 0 | 760 | 1426 |

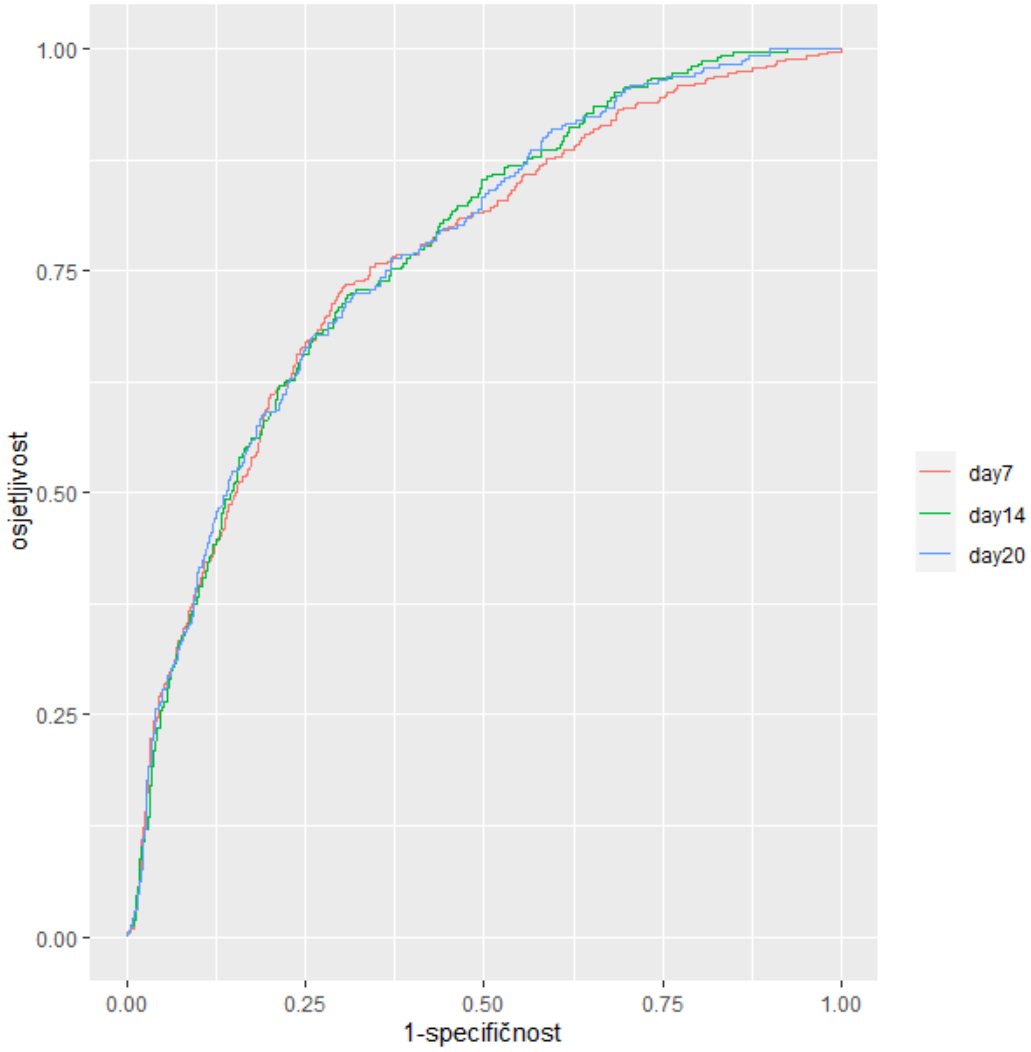
Tablica 4.4: Kontigencijska tablica za *full-model* s interakcijama.

provodimo stepwise proceduru čiji rezultirajući model nazivamo finalni model. Konačno, procjenjujemo finalni model s interakcijama. Dok smo se kod definicije odljeva korisnika $\text{Def}_1$ koncentrirali na izgradnju modela, navodili sve korake stepwise procedure i logističke regresije kao klasifikatora, u preostala tri slučaja koncentrirat ćemo se na promatranje rezultata na skupu za testiranje. U tablici 4.5 prikazani su rezultati finalnih modela za definicije odljeva korisnika $\text{Def}_7$, $\text{Def}_{14}$ i $\text{Def}_{20}$. Napomenimo da smo prilikom odabira granične vrijednosti $c$ za klasifikaciju je li igrač churner ili ne odabrali upravo najčešće korištenu vrijednost 0.5. Nadalje, kao mjeru valjanosti možemo promatrati i točnost jer smo prije same procjene modela izvršili proces balansiranja churnera i ne-churnera. Upravo iz ovog razloga proučili smo udio churnera i ne-churnera na slici 4.1.

| | $\text{Def}_7$ | $\text{Def}_{14}$ | $\text{Def}_{20}$ |
|---|---|---|---|
| AUC | 0.7598 | 0.7674 | 0.7669 |
| osjetljivost | 0.7576 | 0.7516 | 0.75 |
| ACC | 0.6659 | 0.6431 | 0.6469 |

Tablica 4.5: Mjere točnosti procijenjenih modela na skupu za testiranje.

Kad bismo temeljem AUC-a ocjenjivali uspješnost procijenjenih modela kao klasifikatora, u sva tri slučaja imali bismo prihvatljivu klasifikaciju. Primijetimo da s porastom perioda neaktivnosti igrača su točnost i osjetljivost skoro pa nepromjenjive.




Slika 4.4: ROC krivulje na skupu za testiranje.

Na slici 4.4 dani su grafički prikazi ROC krivulja za definicije odljeva korisnika $\text{Def}_7$, $\text{Def}_{14}$ i $\text{Def}_{20}$ što je označeno s *day7*, *day14* i *day20* redom.

# Zaključak

Problem kojim smo se bavili u ovom diplomskom radu je predikcija odljeva korisnika mobilne igre. U svrhu predikcije odljeva korisnika koristi se i razvijeno je mnogo alata, a logistička regresija je jedna od najpopularnijih metoda. Cilj rada bio je prikazati primjenu logističke regresije, a prije same primjene, u radu je uveden koncept logističke regresije te je prikazana osnovna teorijska podloga. Primjena je provedena na setu realnih podataka. Na temelju varijabli konstruiranih iz interakcije korisnika s igrom procijenjeni su multivarijatni prediktivni modeli za različite definicije odljeva korisnika. Modeli su procijenjeni na skupu za treniranje i pri procjeni je korištena metoda za balansiranje podataka.

# Literatura


[1] David W. Hosmer, Stanley Lemeshow (2000) Applied Logistic Regression, John Wiley  Sons, 2000.

[2] Jiawei Han, Micheline Kamber, Jian Pei (2012) Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012.

[3] N. Sarapa (2002) Teorija vjerojatnosti. Zagreb: Skolska knjiga.

[4] Perišić, A., Pahor, M. (2020) Extended RFM logit model for churn prediction in the mobile gaming market. Croatian Operational Research Review 11 (2). doi:10.17535/crorr.2020.0020

TEMELJNA DOKUMENTACIJSKA KARTICA

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU
ODJEL ZA MATEMATIKU

DIPLOMSKI RAD

# LOGISTIČKA REGRESIJA U PREDIKCIJI ODLJEVA KORISNIKA MOBILNE IGRE

Anđela Šarlija

**Sažetak**:

*Logistička regresija je statistička metoda koja se koristi u analizi podataka u kojoj je cilj opisati vezu između dihotomne zavisne varijable i jedne (ili više) nezavisnih varijabli. U ovome radu prikazana je primjena logističke regresije u modeliranju problema odljeva korisnika u industriji mobilnih igara. Nakon prikaza osnovne teorijske podloge, izgrađeni su univarijatni i multivarijatni logistički modeli za problem predikcije odljeva korisnika pri čemu su uspoređeni modeli s obzirom na različite definicije odljeva korisnika. Prikazana je izgradnja multivarijatnih modela temeljem stepwise procedure uz podršku statističkog programa R.*

**Ključne riječi**:

*logit, stepwise metoda, prediktivno modeliranje*

**Podatci o radu**:

*broj stranica 42, broj slika 4, broj tablica 7, broj literaturnih navoda 4, jezik izvornika: hrvatski*

**Mentorica**: *doc. dr. sc. Tea Martinić Bilać*

**Neposredna voditeljica**: dr. sc. Ana Perišić

**Član povjerenstva**:

TEMELJNA DOKUMENTACIJSKA KARTICA

*doc. dr. sc. Aljoša Šubašić*

Povjerenstvo za diplomski rad je prihvatilo ovaj rad *23. rujna 2022.*

TEMELJNA DOKUMENTACIJSKA KARTICA

FACULTY OF SCIENCE, UNIVERSITY OF SPLIT
DEPARTMENT OF MATHEMATICS

MASTER'S THESIS

# LOGISTIC CHURN PREDICTION MODELS FOR MOBILE GAMES

Anđela Šarlija

**Abstract**:

*Logistic regression is a statistical method used in data analysis where the goal is to describe the relationship between a dichotomous dependent variable and one or more independent variables. This paper presents the application of logistic regression in modeling churn problems in the mobile game industry. This work provides the presentation of the basic theoretical basis and the application based on the real-world data where univariate and multivariate logistic models are built for the problem of predicting user churn. Also, models are compared at different levels of churn definition. The construction of multivariate models based on the stepwise method is presented with the support of the statistical program R.*

**Key words**:

*logit, stepwise method, predictive modeling*

**Specifications**:

*42 pages, 4 figures, 7 tables, 4 references, original in: Croatian*

**Mentor**: *assistant professor Tea Martinić Bilać*

**Immediate mentor**: *Ana Perišić, PhD*

TEMELJNA DOKUMENTACIJSKA KARTICA

**Committee**:

*assistant professor Aljoša Šubašić*

This thesis was approved by a Thesis commettee on *September 23, 2022.*