Analiza korištenja mobilne igre analizom doživljenja

Rajković, Lara

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Lara Rajković

ANALIZA KORIŠTENJA MOBILNE IGRE ANALIZOM DOŽIVLJENJA

Diplomski rad

Voditelj rada:
prof. dr. sc.
Anamarija Jazbec

Zagreb, rujan 2020.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Analiza doživljenja odnosi se na statističke metode u kojima je uz promatranu varijablu događaja analizirana varijabla vremena do tog događaja. Metode analize doživljenja prilagođene su tome što vrijeme do događaja najčešće nije normalno distribuirano i događaj najčešće nije opažen kod svih subjekata.

Začeci analize doživljenja potječu još iz 17. stoljeća, ali tri najbitnije publikacije su Kaplan i Meier iz 1958. [3] koji su razvili neparametarsku procjenu funkcije doživljenja, David Cox, 1972. [2] koji je razvio Coxovu regresiju i rad Odd Aalena [1] koji je razvio teoriju brojećih procesa i njihovu primjenu u analizi doživljenja, počevši 1975.

Analizom doživljenja originalno se promatralo vrijeme do smrti od početka tretmana pojedinca, otkud joj potječe i ime, ali se s vremenom primjenjuje u raznim disciplinama kao što su aktuarstvo, biomedicina, inženjerstvo, sociologija itd. U ovom radu primijenjene su metode analize doživljenja na skup podataka iz mobilne igre gdje je promatrani događaj prestanak igranja korisnika (na engleskom *churn*). Primijenjene su neparametrijske metode kao što je Kaplan Meierova procjena te semi-parametrijski Coxov regresijski model. U prvom poglavlju je kratki pregled teorije analize doživljenja, a u drugom primjena navedenih metoda na podacima.

Poglavlje 1

Analiza doživljenja

1.1 Osnovne statističke metode

Funkcija doživljenja

Neka je X nenegativna slučajna varijabla koja predstavlja vrijeme do nekog specifičnog promatranog događaja.

Funkciju $F : [0, +\infty\rangle \to [0, 1]$ definiranu s:

$$F(x) = \mathbb{P}(X \leq x)$$

zovemo funkcija distribucije slučajne varijable X.

Za nju vrijede svojstva:

$$F(0) = 0 \text{ i } \lim_{x\to\infty} F(x) = 1.$$

Definiramo funkciju doživljenja $S : [0, +\infty\rangle \to [0, 1]$ slučajne varijable X:

$$S(x) = \mathbb{P}(X > x).$$

Ona predstavlja vjerojatnost pojedinca da doživi događaj nakon vremena x.

Za funkciju doživljenja vrijedi $S(x) = 1 - F(x)$ gdje je F funkcija distribucije slučajne varijable X. Iz svojstva funkcije distribucije slučajne varijable X slijedi da je S padajuća funkcija za koju vrijedi $S(0) = 1$ i $\lim_{x\to\infty} S(x) = 0$.

Ako je X neprekidna slučajna varijabla, postoji nenegativna funkcija f za koju vrijedi

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} f(t)dt, \quad \text{za sve } x \geq 0$$

i ona se naziva funkcija gustoće slučajne varijable X. Za funkciju doživljenja vrijedi

$$S(x) = \int_{x}^{\infty} f(t)dt, \quad x \geq 0$$

i za funkciju gustoće vrijedi

$$f(x) = -S'(x).$$

Varijabla X može biti i diskretna slučajna varijabla i to se pojavljuje u praksi zbog zaokruživanja mjernih jedinica, grupiranja vremena doživljenja i slično. Neka X poprima vrijednosti x_j, $j \in \mathbb{N}$ s vjerojatnostima p_j, $j \in \mathbb{N}$, tada je funkcija doživljenja dana s

$$S(x) = \sum_{\substack{j \in \mathbb{N} \\ x_j > x}} p_j.$$

Funkcija hazarda i kumulativna funkcija hazarda

Funkcija hazarda $h : [0, +\infty\rangle \to [0, +\infty\rangle$ za neprekidnu slučajnu varijablu X definirana je s

$$h(x) = \lim_{\Delta x \to 0} \frac{\mathbb{P}(x \leq X < x + \Delta x \mid X \geq x)}{\Delta x}.$$

Iz definicije vidimo da $h(x)\Delta x$ možemo promatrati kao približnu vjerojatnost pojedinca koji je doživio vrijeme x da doživi događaj u sljedećem trenutku. Funkcijom hazarda možemo opisati kako se vjerojatnost događaja mijenja kroz vrijeme. Jedina restrikcija na h je da mora biti nenegativna. Funkcija može biti raznih oblika, ovisno o promatranom događaju. Ovisno o tome je li funkcija hazarda konstantna, rastuća, padajuća, „oblika grbe" ili „oblika kade", možemo odrediti matematičku funkciju koja opisuje funkciju hazarda. Tako modelirani modeli doživljenja su parametrijski, s pripadnom parametrijskom funkcijom doživljenja. Neki od parametrijskih modela doživljenja su eksponencijalni, Weibull, Gompertz-Makeham, gama, log-normalni itd.

Povežimo funkciju hazarda s funkcijom doživljenja.

$$\mathbb{P}(X > x + \Delta x \mid X \geq x) = \frac{S(x + \Delta x)}{S(x)}$$

slijedi da je

$$h(x) = -\frac{S'(x)}{S(x)} = \frac{f(x)}{S(x)} = -\frac{d}{dx}\left(ln\left(S(x)\right)\right).$$

Kumulativna funkcija hazarda $H\,[0, +\infty\rangle \to [0, +\infty\rangle$ definirana je s

$$H(x) = \int_0^x h(u)du$$

i ona je blisko povezana s funkcijom doživljenja relacijama:

$$H(x) = -ln\left(S(x)\right), \quad S(x) = e^{-H(x)}.$$

Ako je X diskretna slučajna varijabla, funkcija hazarda dana je s

$$h(x_j) = \mathbb{P}\Big(X = x_j \mid X \geq x_j\Big) = \frac{p_j}{S(x_{j-1})}, \quad j \in \mathbb{N}$$

pri čemu je $S(x_0) = 1$.

Funkciju doživljenja možemo zapisati kao

$$S(x) = \prod_{x_j \leq x} \frac{S(x_j)}{S(x_{j-1})}$$

pa su funkcija doživljenja i funkcija hazarda povezane s

$$S(x) = \prod_{x_j \leq x} \Big[1 - h(x_j)\Big].$$

Kumulativnu funkciju hazarda za diskretnu varijablu vremena definiramo s

$$H(x) = -\sum_{x_j \leq x} ln\Big[1 - h(x_j)\Big]$$

pa relacije $H(x) = -ln\,(S(x))$ i $S(x) = e^{-H(x)}$ i u ovom slučaju vrijede.

Cenzuriranje

Podaci opažanja vremena do događaja specifični su po tome što vrijeme događaja subjekta nije uvijek poznato. Opažanja u kojima vrijeme događaja nije poznato nazivamo cenzurirana opažanja. Opažanja mogu biti desno, lijevo ili intervalno cenzurirana. Za desno cenzurirana opažanja znamo da je vrijeme događaja subjekta veće od promatranog vremena tj. od vremena cenzuriranja. Razlog tome je najčešće to što ne dožive svi subjekti promatrani događaj za vrijeme trajanja eksperimenta ili što napuste eksperiment prijevremeno. Za lijevo cenzurirana opažanja znamo da je vrijeme događaja manje od vremena cenzuriranja tj. događaj se dogodio prije nego što je subjekt promatran. Intervalno cenzurirana opažanja su ona za koja znamo samo da se događaj dogodio u nekom vremenskom intervalu. Desno cenzuriranje je najčešće i rezultati koji su ovdje navedeni vrijede za skupove podataka s desno cenzuriranim opažanjima. Desno cenzurirana opažanja nastaju zbog unaprijed određene duljine trajanja eksperimenta ili zbog unaprijed određenog broja jedinki koje dožive događaj. Također nastaje zbog drugih nepredvidivih faktora zbog kojih više ne možemo promatrati subjekt.

Pretpostavimo da imamo n subjekata u eksperimentu. Za svaki subjekt neka je $X_i, i = 1, \ldots, n$ vrijeme doživljenja tj. vrijeme do promatranog događaja. Vremena X_i su nezavisna i jednakodistribuirana s funkcijom gustoće f i funkcijom doživljenja S. Neka je C_i

vrijeme cenzuriranja subjekta i. Desno cenzuriranje *tipa I* je ono koje nastaje zbog unaprijed određene duljine eksperimenta. U tom slučaju je vrijeme $C_i, i = 1, \dots n$ unaprijed određeno. Kod desnog cenzuriranja *tipa II* eksperiment traje dok r subjekata ne doživi događaj, pri čemu je r unaprijed određen broj, $r < n$. Vrijeme cenzuriranja je tada $C_i = T_{(r)}$ i ono je slučajno. Još jedna vrsta desnog cenzuriranja je slučajno cenzuriranje. Ono nastaje pod utjecajem vanjskih faktora zbog kojih subjekt prestaje biti u eksperimentu i više ne možemo pratiti njegovo vrijeme događaja. U ovom slučaju je $C_i, i = 1, \dots, n$ potencijalno vrijeme cenzuriranja.

Točno vrijeme doživljenja subjekta i znamo ako i samo ako je $X_i \leq C_i$. U suprotnom je vrijeme događaja cenzurirano u vremenu C_i. Slučajni uzorak iz ovakvog eksperimenta prikazujemo pomoću uređenih parova slučajnih varijabli $(T_i, \delta_i), i = 1 \dots n$ pri čemu je $T_i = \min(X_i, C_i)$, a δ_i je indikator događaja za i-tog subjekta tj.

$$\delta_i = \begin{cases} 0, & \text{za } T_i = C_i \\ 1, & \text{za } T_i = X_i. \end{cases}$$

Kaplan-Meier procjenitelj i Nelson-Aalen procjenitelj

Neka je $(t_1, \delta_1) \dots (t_n, \delta_n)$ opaženi uzorak s cenzuriranim događajima. Pretpostavimo da su vremena cenzuriranja nezavisna s vremenima događaja. Pretpostavimo da su se događaji dogodili u D različitih vremena $t_{(1)} < t_{(2)} < \cdots < t_{(D)}$ i da se u vremenu $t_{(i)}$ dogodilo d_i događaja. Neka je Y_i broj subjekata kojima je vrijeme događaja veće ili jednako od $t_{(i)}$. Za takve subjekte kažemo da su pod rizikom u vremenu $t_{(i)}$.

Kaplan-Meierov procjenitelj funkcije doživljenja je definiran s

$$\hat{S}(t) = \begin{cases} 1 & \text{za } t < t_{(1)} \\ \prod_{t_{(i)} \leq t} \left[1 - \frac{d_i}{Y_i}\right] & \text{za } t_{(1)} \leq t. \end{cases}$$

Procjenitelj je definiran do vrijednosti najvećeg opaženog vremena. Tako dobiven procjenitelj je step-funkcija koja ima „skokove" kod opaženih vremena događaja. Veličine „skokova" ne ovise samo o broju događaja nego i o broju cenzuriranih opažanja. Vrijednost dobivene funkcije u vremenu t naziva se i vjerojatnost doživljenja u vremenu t.

Varijanca procjenitelja dana je Greenwoodovom formulom:

$$\hat{V}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_{(i)} \leq t} \frac{d_i}{Y_i (Y_i - d_i)}.$$

Kumulativnu funkciju hazarda možemo procijeniti pomoću funkcije doživljenja kao

$$\hat{H}(t) = -ln\left[\hat{S}(t)\right].$$

Alternativni procjenitelj kumulativne funkcije hazarda koji je bolji za manje uzorke je Nelson-Aalenov procjenitelj koji je definiran do vrijednosti najvećeg opaženog vremena:

$$\hat{H}(t) = \begin{cases} 0 & \text{za } t < t_{(1)} \\ \sum_{t_{(i)} \le t} \frac{d_i}{Y_i} & \text{za } t_{(1)} \le t \end{cases}$$

i njegova varijanca dana je s:

$$\hat{V}(\hat{H}(t)) = \sum_{t_{(i)} \le t} \frac{d_i}{Y_i^2}.$$

Funkciju doživljenja tada možemo procijeniti s

$$\hat{S}(t) = e^{-\hat{H}(t)}.$$

Testovi za usporedbu funkcija doživljenja

Želimo usporediti funkcije doživljenja K populacija tj. testirati sljedeću hipotezu

$$H_0 : S_1(t) = S_2(t) = \cdots = S_K(t), \text{ za sve } t \le \tau,$$
$$H_1 : \text{barem jedna od } S_j(t) \text{ je različita za neki } t \le \tau.$$

Ovdje je τ najveće vrijeme u kojem sve grupe imaju barem jednu jedinku pod rizikom. Neka su $t_{(1)} < t_{(2)} < \cdots < t_{(D)}$ različita vremena događaja u zajedničkom uzorku. U vremenu $t_{(i)}$ opaženo je d_{ij} događaja u j-tom uzorku od Y_{ij} subjekata pod rizikom, $j = 1, 2, \ldots, K, i = 1, 2, \ldots, D$. Neka je $d_i = \sum_{j=1}^{K} d_{ij}$ broj događaja i $Y_i = \sum_{j=1}^{K} Y_{ij}$ u kombiniranom uzorku u vremenu $t_{(i)}$, $i = 1, \ldots, D$.

Test se bazira na statistici

$$Z_j(\tau) = \sum_{i=1}^{D} W(t_{(i)}) \left[d_{ij} - Y_{ij} \left(\frac{d_i}{Y_i} \right) \right], \quad j = 1, \ldots, K$$

gdje je W težinska funkcija. Testna statistika je suma težinskih razlika između opaženog broja događaja i očekivanog broja događaja u j-tom uzorku pod pretpostavkom da H_0 vrijedi.

Varijanca od $Z_j(\tau)$ je dana s

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} W(t_{(i)})^2 \frac{Y_{ij}}{Y_i} \left(1 - \frac{Y_{ij}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i, \quad j = 1, \ldots, K$$

i kovarijanca od $Z_j(\tau)Z_g(\tau)$ je dana s

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W(t_{(i)})^2 \frac{Y_{ij}}{Y_i}\frac{Y_{ig}}{Y_i}\left(\frac{Y_i - d_i}{Y_i - 1}\right)d_i, \quad g \neq j.$$

Procijenjena matrica varijanci i kovarijanci sastavljena od $\hat{\sigma}_{jg}$ je dana s $(K-1)\times(K-1)$ matricom Σ. Testna statistika za testiranje hipoteze je

$$(Z_1(\tau), \ldots, Z_{K-1}(\tau))\Sigma^{-1}(Z_1(\tau), \ldots, Z_{K-1}(\tau))^t \sim \chi^2(K-1).$$

Postoji više testova koji se razlikuju u izboru funkcije W. Najčešće se primjenjuje *log-rank* test u kojem je $W(t) = 1$.

1.2 Coxov regresijski model

Model

Često za subjekte postoje neke karakteristike za koje nas zanima utječu li na njihov ishod u eksperimentu. Te varijable možemo koristiti kao prediktore tj. nezavisne varijable u modelu koje objašnjavaju zavisnu varijablu koja nosi informaciju je li se događaj dogodio i kada. Također dobiveni model možemo koristiti za predviđanje ishoda i vremena događaja subjekta s određenim karakteristikama.

Sada se uzorak sastoji od uređenih trojki (T_j, δ_j, Z_j), $j = 1, 2, \ldots, n$ gdje je T_j vrijeme događaja ili cenzuriranja za j-tog subjekta, δ_j indikatorska varijabla i $Z_j = (Z_{j1}, \ldots, Z_{jp})^t$ vektor nezavisnih varijabli za j-tog subjekta u vremenu t. Neka je $h(t \mid \mathbf{Z})$ funkcija hazarda u vremenu t za subjekta s vektorom varijabli $\mathbf{Z}$. Osnovni Coxov model je:

$$h(t \mid \mathbf{Z}) = h_0(t)c(\boldsymbol{\beta}^t\mathbf{Z})$$

gdje je h_0 bazna funkcija hazarda, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ je vektor koeficijenata i c je poznata funkcija. Model se naziva semi-parametrijski jer distribucija bazne funkcije nije poznata, ona se tretira neparametrijski, dok je efekt nezavisnih varijabli parametrijski.

Najčešće se uzima da je

$$c(\boldsymbol{\beta}^t\mathbf{Z}) = \exp(\boldsymbol{\beta}^t\mathbf{Z}) = \exp(\sum_{k=1}^{p}\beta_k Z_k)$$

pa je Coxov regresijski model

$$h(t \mid \mathbf{Z}) = h_0(t)\exp(\boldsymbol{\beta}^t\mathbf{Z}) = h_0(t)\exp(\sum_{k=1}^{p}\beta_k Z_k).$$

Coxov model naziva se model proporcionalnog hazarda jer za subjekte koji imaju vrijednosti kovarijata $\mathbf{Z}$ i $\mathbf{Z}^*$ njihov omjer funkcija hazarda je konstantan i označava se s *HR* (*hazard ratio*):

$$HR = \frac{h(t \mid \mathbf{Z})}{h(t \mid \mathbf{Z}^*)} = \frac{h_0(t)\exp(\sum_{k=1}^{p}\beta_k Z_k)}{h_0(t)\exp(\sum_{k=1}^{p}\beta_k Z_k^*)} = \exp\left(\sum_{k=1}^{p}\beta_k(Z_k - Z_k^*)\right)$$

Procjena parametara i odabir varijabli za model

Pretpostavimo da su vremena događaja različita za svaki promatrani subjekt i da je vrijeme događaja nezavisno s vremenom cenzuriranja za svaki subjekt. Neka su $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$ uređena vremena događaja i $Z_{(i)k}$ je k-ti vektor prediktora subjekta čije je vrijeme doživljenja $t_{(i)}$. Neka je $R(t_{(i)})$ skup subjekata pod rizikom u vremenu $t_{(i)}$. Funkcija parcijalne vjerodostojnosti dana je s

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(\sum_{k=1}^{p}\beta_k(Z_{(i)k}\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\sum_{k=1}^{p}\beta_k(Z_{jk}\right)}.$$

Procjenitelj β se dobiva logaritmirajući funkciju parcijalne vjerodostojnosti i tražeći njezin maksimum. Dakle, dobivamo:

$$LL(\boldsymbol{\beta}) = ln(L(\beta)) = \sum_{i=1}^{D}\sum_{k=1}^{p}\beta_k Z_{(i)k} - \sum_{i=1}^{D} ln\left(\sum_{j\in R(t_{(i)})}\exp\left(\sum_{k=1}^{p}\beta_k Z_{jk}\right)\right),$$

$$U_h(\boldsymbol{\beta}) = \frac{\delta LL(\beta)}{\delta\beta_h} = \sum_{i=1}^{D} Z_{(i)h} - \sum_{i=1}^{D}\frac{\sum_{j\in R(t_{(i)})} Z_{jh}\exp\left(\beta_k Z_{jk}\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\beta_k Z_{jk}\right)}, \quad h = 1, \ldots, p$$

pa se procjenitelji maksimalne parcijalne vjerodostojnosti dobiju iz $U_h(\beta) = 0, h = 1, .., p$ pomoću neke od iterativnih metoda.

Često u praksi više subjekata ima jednako vrijeme doživljenja. U tom slučaju, postoji više predloženih funkcija parcijalne vjerodostojnosti.

Neka su $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$ različita, uređena vremena događaja. Neka je d_i broj događaja u vremenu $t_{(i)}$ i $\mathbb{D}_i$ skup subjekata koji imaju događaj u vremenu $t_{(i)}$. Neka je $\boldsymbol{s_i} = \sum_{j\in\mathbb{D}_i}\boldsymbol{Z_j}$ i R_i skup subjekata pod rizikom netom prije $t_{(i)}$.

Jedna od najčešće korištenih funkcija parcijalne vjerodostojnosti u tom slučaju je Breslowova (1974):

$$L_B(\boldsymbol{\beta}) = \prod_{i=1}^{D}\frac{\exp\left(\boldsymbol{\beta^t s_i}\right)}{\left[\sum_{j\in R_i}\exp\left(\boldsymbol{\beta^t Z}_j\right)\right]^{d_i}}.$$

Ova aproksimacija je dobra kada nema puno jednakih vremena događaja.
Efron (1977) je predložio sljedeću funkciju parcijalne vjerodostojnosti:

$$L_E(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{\beta^t s_i})}{\prod_{j=1}^{d_i}\left[\sum_{k\in R_i}\exp(\boldsymbol{\beta^t Z_k}) - \frac{j-1}{d_i}\sum_{k\in\mathbb{D}_i}\exp(\boldsymbol{\beta^t Z_k})\right]},$$

koja je bliže egzaktnoj funkciji parcijalne vjerodostojnosti temeljene na diskretnom modelu. Ako je broj jednakih vremena subjekata mali, ova dva procjenitelja daju gotovo isti rezultat.

Egzaktna funkcija parcijalne vjerodostojnosti temeljena na diskretnom modelu pogodna je za modeliranje diskretne varijable vremena, ali za veliku količinu podataka zahtjevna je za računanje, čak i za današnja računala.

Za testiranje parametra modela najviše se koriste tri testa: Waldov test, test omjera vjerodostojnosti i test skorova (engleski *score test*). Neka je $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots \hat{\beta}_p)^t$ procjenitelj za $\boldsymbol{\beta}$ dobiven pomoću parcijalne maksimalne vjerodostojnosti, $I(\boldsymbol{\beta})$ je $p \times p$ informacijska matrica koja se računa kao $I_{gh}(\boldsymbol{\beta}) = -\frac{\delta^2 LL(\beta)}{\delta\beta_g\delta\beta_h}, g = 1, \ldots, p, h = 1, \ldots, p$ i $\boldsymbol{U(\beta)} = (U_1(\boldsymbol{\beta}), \ldots, U_p(\boldsymbol{\beta}))^t$. Želimo testirati hipotezu

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}.$$

Testne statistike navedenih testova su redom:

$$\begin{aligned}
\chi^2_W &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})^t \boldsymbol{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}),\\
\chi^2_{LR} &= 2\left(LL(\hat{\boldsymbol{\beta}}) - LL(\boldsymbol{\beta_0})\right),\\
\chi^2_{SC} &= \boldsymbol{U}(\boldsymbol{\beta_0})^t \boldsymbol{I}^{-1}(\boldsymbol{\beta_0})\boldsymbol{U}(\boldsymbol{\beta_0}).
\end{aligned}$$

Sve tri testne statistike pod uvjetom da H_0 vrijedi imaju χ^2 distribuciju s p stupnjeva slobode za velike uzorke.

Za testiranje pojedinih parametara modela $\beta_i, i = 1, \ldots, p$ također se koriste navedeni testovi, npr. za testiranje hipoteze $H_0 : \beta_1 = 0$ Waldova testna statistika je $\frac{\hat{\beta}_1^2}{Var(\hat{\beta}_1)} \sim \chi^2(1)$ pri čemu se procjenitelj za $Var(\hat{\beta}_1)$ može dobiti pomoću informacijske matrice.

Za usporedbu ugnježđenih modela, kao i za ostale regresijske modele, koristi se test omjera vjerodostojnosti. Neka jednostavniji model sadrži varijable $Z_1, Z_2, \ldots, Z_p$ i procjenitelj dobiven metodom parcijalne maksimalne vjerodostojnosti je $\hat{\boldsymbol{\beta}}_\mathbf{1}$, a prošireni varijable $Z_1, Z_2, \ldots, Z_p, Z_{p+1} \ldots, Z_{p+q}$ s procjeniteljem $\hat{\boldsymbol{\beta}}_\mathbf{2}$ i želimo testirati hipotezu

$$\begin{aligned}
&H_0 : \beta_{p+1} = \beta_{p+2} = \ldots = \beta_{p+q} = 0\\
&H_1 : \beta_j \neq 0, \text{za neki} j \in \{p+1, \ldots, p+q\}.
\end{aligned}$$

Testna statistika je

$$2\left(LL(\hat{\boldsymbol{\beta}}_\mathbf{2}) - LL(\hat{\boldsymbol{\beta}}_\mathbf{1})\right) \sim \chi^2(q).$$

Procjena funkcije doživljenja i hazarda pomoću modela

Pomoću dobivenog modela možemo procijeniti funkciju doživljenja subjekta s određenim karakteristikama. Neka je

$$W(t_{(i)};\hat{\boldsymbol{\beta}}) = \sum_{j\in R(t_i)} \exp\left(\sum_{h=1}^{p} \hat{\beta}_h Z_{jh}\right).$$

Breslowov procjenitelj bazične kumulativne funkcije hazarda $H_0(t) = \int_0^t h_0(u)du$ dan je s

$$\hat{H}_0(t) = \sum_{t_{(i)}\le t} \frac{d_i}{W(t_i;\hat{\boldsymbol{\beta}})}$$

i jednak je Nelson-Aalenovom procjenitelju kad nema prisutnih kovarijata u modelu.

Bazična funkcija doživljenja sada se procjenjuje s

$$\hat{S}_0(t) = \exp\left(-\hat{H}_0(t)\right)$$

i to je procijenjena funkcija doživljenja za subjekta kojem su sve kovarijate jednake 0. Za subjekta s kovarijatama $\boldsymbol{Z = Z_0}$ procijenjena funkcija doživljenja je

$$\hat{S}(t \mid \boldsymbol{Z = Z_0}) = \hat{S}_0(t)^{\exp(\hat{\boldsymbol{\beta}}^t \mathbf{Z_0})}.$$

Združivanje i interakcije

Coxov regresijski model može biti univarijatni i multivarijatni. U praksi se nekad model izgrađuje kako bi se saznao utjecaj samo jednog, glavnog prediktora, a nekad više njih. Pri izgradnji modela treba paziti na tzv. združivanje do kojeg dolazi zbog međusobne zavisnosti prediktora. Ako se koeficijent uz prediktor, u modelu u kojem je samo jedan prediktor, jako razlikuje od koeficijenta uz prediktor u modelu u kojem je i druga varijabla, to nam ukazuje da su te varijable međusobno zavisne i da je došlo do združivanja. U tom slučaju obje varijable trebaju biti u modelu.

Također, jedna varijabla može mijenjati efekt druge varijable. Tada kažemo da postoji interakcija između te dvije varijable. Neka je dan model s dvije nezavisne varijable

$$h(t \mid (Z_1, Z_2)) = h_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2).$$

Ako postoji interkacija između njih, tada model postaje

$$\begin{aligned} h(t \mid (Z_1, Z_2)) &= h_0(t)\exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2) \\ &= h_0(t)\exp((\beta_1 + \beta_3 Z_2)\, Z_1 + \beta_2 Z_2) \\ &= h_0(t)\exp((\beta_2 + \beta_3 Z_1)\, Z_2 + \beta_1 Z_1) \end{aligned}$$

pri čemu je $\beta_3 \neq 0$. Vidimo da ako je interkacija prisutna, efekt varijable Z_1 mijenja se s promjenom vrijednosti varijable Z_2 i obrnuto.

Varijable ovisne o vremenu

Pomoću Coxovog regresijskog modela možemo mjeriti i efekt varijabli koje se mijenjaju kroz vrijeme u eksperimentu. Takve varijable mogu biti diskretne i neprekidne. Pretpostavka je da znamo vrijednost varijable u svakom vremenu t u kojem je subjekt promatran. Sada je vektor nezavisnih varijabla jednak $Z_j(t) = (Z_{j1}(t), \ldots, Z_{jp}(t))^t$ gdje varijable mogu biti fiskne ili promjenjive u vremenu te Coxov regresijski model zapisujemo kao

$$h(t \mid \mathbf{Z}(t)) = h_0(t)\exp(\boldsymbol{\beta}^t\mathbf{Z}(t)) = h_0(t)\exp(\sum_{k=1}^{p}\beta_k Z_k(t)).$$

Provjera pretpostavke proporcionalnog hazarda

Osnovna pretpostavka Coxovog regresijskog modela je proporcionalnost hazarda - u svakom vremenu u eksperimentu je omjer funkcija hazarda različitih subjekata jednak. To u praksi ne mora vrijediti i više je načina da se ta pretpostavka provjeri. Jedan od načina provjere pretpostavke je pomoću varijabli ovisnih o vremenu. Za provjeru pretpostavke za fisknu varijablu Z_1, uvede se varijabla ovisna o vremenu kao interakcija između fiksne varijable i neke funkcije ovisne o vremenu:

$$Z_2(t) = Z_1 \times g(t).$$

Najčešće se uzima $g(t) = ln(t)$. Zatim se prilagodi Coxov model s varijablama Z_1 i $Z_2(t)$ s koeficijentima β_1 i β_2 pa je funckija hazarda u vremenu t jednaka

$$h(t \mid Z_1) = h_0(t)\exp\left(\beta_1 Z_1 + \beta_2 \left(Z_1 \times g(t)\right)\right)$$

i omjer hazarda dvaju subjekata je

$$\frac{h(t \mid Z_1)}{h(t \mid Z_1^*)} = \exp\left(\beta_1 \left(Z_1 - Z_1^*\right) + \beta_2 g(t) \left(Z_1 - Z_1^*\right)\right).$$

Taj omjer ovisi o t ako je $\beta_2 \neq 0$ pa za testiranje pretpostavke proporcionalnog hazarda zapravo testiramo hipotezu $H_0 : \beta_2 = 0$.

U slučaju da pretpostavka proporcionalnog hazarda nije zadovoljena i promatrana varijabla Z_1 je dihotomna, jedno od rješenja je uvesti varijable ovisne o vremenu Z_2 i Z_3 definirane s

$$Z_2(t) = \begin{cases} Z_1 & \text{za } t \leq \tau \\ 0 & \text{za } t > \tau, \end{cases} \qquad Z_3(t) = \begin{cases} 0 & \text{za } t \leq \tau \\ Z_1 & \text{za } t > \tau. \end{cases}$$

Na taj način dobivamo model

$$h(t \mid Z(t)) = \begin{cases} h_0(t)\exp(\beta_1 Z_1) & \text{za } t \leq \tau \\ h_0(t)\exp(\beta_2 Z_1) & \text{za } t > \tau \end{cases}$$

u kojem je koeficijent uz varijablu Z_1 ovisan o vremenu. Za izbor vremena τ možemo napraviti modele s različitim vrijednostima τ i zatim izabrati model s najvećom vrijednosti logaritmirane funkcije parcijalne vjerodostojnosti.

Poglavlje 2

Primjena analize doživljenja na podacima iz mobilne igre

2.1 Podaci

Podaci su uzeti iz jedne *free-to-play* mobilne igre. Igra je besplatna za instaliranje i igranje, ali korisnik može kupovati valutu koja se koristi u igri kako bi brže napredovao ili imao stvari koje mu nisu dostupne ako ne plati. Za takve igre je bitno da korisnici što dulje igraju kako bi što više novaca potrošili u igri. U takvim igrama puno korisnika odustane od igre nakon jednog dana, ali dio korisnika igra čak i godinama. Ne postoji trenutak za koji točno možemo reći da je korisnik odustao od igre i da se više neće vratiti. U ovom radu uzeto je da je korisnik odustao ako nije otvorio igru 10 dana. To naravno ne vrijedi za sve i neki se vraćaju nakon 10 dana, ali ako korisnik ne igra 10 dana možemo reći da je izgubio interes za igru. Također ne možemo uzeti dovoljno veliki period u kojem su svi promatrani korisnici odustali od igre jer neki igraju godinama. Zato dolazi do cenzuriranih opažanja. U ovom radu promatramo 41 dan od instalacije igre.

Događaj koji je promatran je *churn*, tj. prestanak igranja korisnika. Uzeti su podaci od korisnika koji su instalirali igru u razdoblju od 8. – 10. svibnja 2020. godine iz 10 država u kojima je bilo najviše instalacija u tom razdoblju. Početak eksperimenta je dan instalacije i kraj eksperimenta je 41 dan nakon instalacije korisnika. Događaj se dogodio tj. korisnik je prestao igrati ako 10 dana nije bio aktivan tj. nije otvorio igru. Ako je korisnik bio zadnji put aktivan od 32. do 41. dana nakon instalacije, a prije toga nije imao vremenski period od 10 dana bez otvaranja igre, on nije doživio događaj te je njegovo opažanje cenzurirano. Varijable vremena su zabilježene u danima jer je za igru predviđeno da je korisnik igra svaki dan, ali je ne igra cijeli dan pa nas ne zanima točno vrijeme odustajanja nego dan od instalacije. Dakle, modelirana varijabla vremena predstavlja redni dan igranja od instalacije, npr. prvih 24 h od instalacije ima vrijednost 1, sljedećih 24 h

ima vrijednost 2 i ako je prošlo 40 punih dana nakon instalacije vrijednost je 41. Uzete su neke kategorijske varijable po kojima se korisnici razlikuju i koje bi mogle utjecati na igranje korisnika. Igra je originalno na engleskom jeziku, a u ovom promatranom vremenu je također dostupna na jezicima: njemački, španjolski, francuski, talijanski i ruski. Postoji više vrsta marketinških kampanja pomoću kojih korisnici instaliraju igru, a neki je sami nađu na trgovini aplikacija i njih nazivamo organski korisnici.

2.2 Metode

Promatrane varijable su:
t događaj - vrijeme u kojem je korisnik odustao ili je cenzuriran,
događaj - ima vrijednost 1 ako je korisnik odustao od igre, 0 ako nije,
mos - mobilni operativni sustav, Android ili iOS,
država - država u kojoj je korisnik instalirao igru, može biti:
br - Brazil,
de - Njemačka,
es - Španjolska,
fr - Francuska,
gb - Velika Britanija,
in - Indija,
it - Italija,
mx - Meksiko,
ru - Rusija,
us - Sjedinjene Američke Države.
kampanja - vrsta marketinške kampanje koja je potaknula korisnika da instalira igru, ako postoji. Može biti:
AEO (*app event optimization*) - kampanja kojoj je cilj prikupiti korisnike koji su slični onima koji imaju neki događaj u igri,
MAI (*mobile app install*) - kampanja kojoj je cilj prikupiti što više korisnika,
VO (*value optimization*) - kampanja kojoj je cilj prikupiti korisnike koji su slični onima koji kupuju sadržaj u igri,
other - ostale kampanje,
organic - korisnik nije instalirao pomoću kampanje.
t kupnja - vrijeme do prve kupnje ako se dogodila, ako nije nema vrijednosti,
platitelj - ima vrijednost 1 ako je korisnik kupio nešto u promatranom vremenu, 0 ako nije, varijabla je ovisna o vremenu.

Za varijable vremena napravljena je opisna statistika s oznakama:
aritmetička sredina - $\bar{x}$,
standardna devijacija - s,

donji kvartil - q_L,
gornji kvartil - q_U,
Za kategorijske varijable napravljena je tablica frekvencija.

Za procjenu funkcije doživljenja korištena je Kaplan-Meierovova procjena, za cijeli uzorak i po kategorijama. *Log-rank* test je primijenjen za usporedbu funkcija doživljenja grupa. Za ispitivanje razlika između svake dvije grupe korišten je *log-rank* test s Bonferronijevom korekcijom *p*-vrijednosti.

Procijenjen je univarijatni Coxov regresijski model s varijablom *mos*. Testirana je pretpostavka proporcionalnosti hazarda uvođenjem interakcije između varijable i logaritma vremenske varijable u model. Model je modificiran uvođenjem koeficijenata ovisnih o vremenu. Zatim je građen multivarijatni model, testirajući koeficijente uz dodane varijable. Varijabla *platitelj* je ovisna o vremenu *t događaj* pa je kod prilagođen tome. Testirana je pretpostavka proporcionalnog hazarda za sve fiksne varijable i model je modificiran. Korišten je test omjera vjerodostojnosti kako bi se ispitalo je li bolji model s interakcijom ili bez nje i tako su dodane statistički značajne interakcije. Za izgradnju funkcije parcijalne vjerodostojnosti kod svih Coxovih regresijskih modela korištena je Effronova metoda.

Za sve statističke testove uzeta je razina značajnosti od 5%. Sve statističke analize i grafički prikazi napravljeni su koristeći programski jezik *R*.

2.3 Rezultati

Opisna statistika

Tablica 2.1: Opisna statistika za varijable *t događaj* i *t kupnja*

varijabla	n	$\bar{x}$	s	min	q_L	medijan	q_U	max
t događaj	47582	4.50	8.72	1	1	1	3	41
t kupnja	1108	4.45	6.15	1	1	2	5	41

Prosječno vrijeme do odustajanja ili cenzuriranja (tablica 2.1) je 4.50 dana, dok je medijan 1 dan što ukazuje na to da je distribucija varijable *t događaj* pozitivno asimetrična. Postotak korisnika koji nisu odustali u promatranom vremenu (tablica 2.5) je 4.32%. Prosječno vrijeme do prve kupnje korsnika koji su platitelji je 4.45 dana, dok je medijan 2 dana, a njih je 2.33% (tablica 2.6). Asimetričnost distribucija varijabli također možemo vidjeti na histogramima relativnih frekvencija. Na slici 2.1 vidimo da je više od 60% korisnika u uzorku odustalo prvi dan igranja. Drugi dan igranja je postotak manji od 10% i dalje se postupno smanjuje. U 41. danu igranja je najveći broj cenzuriranih opažanja, to

Tablica 2.2: Opisna statistika za varijable grupirane po kategorijama varijable *mos*

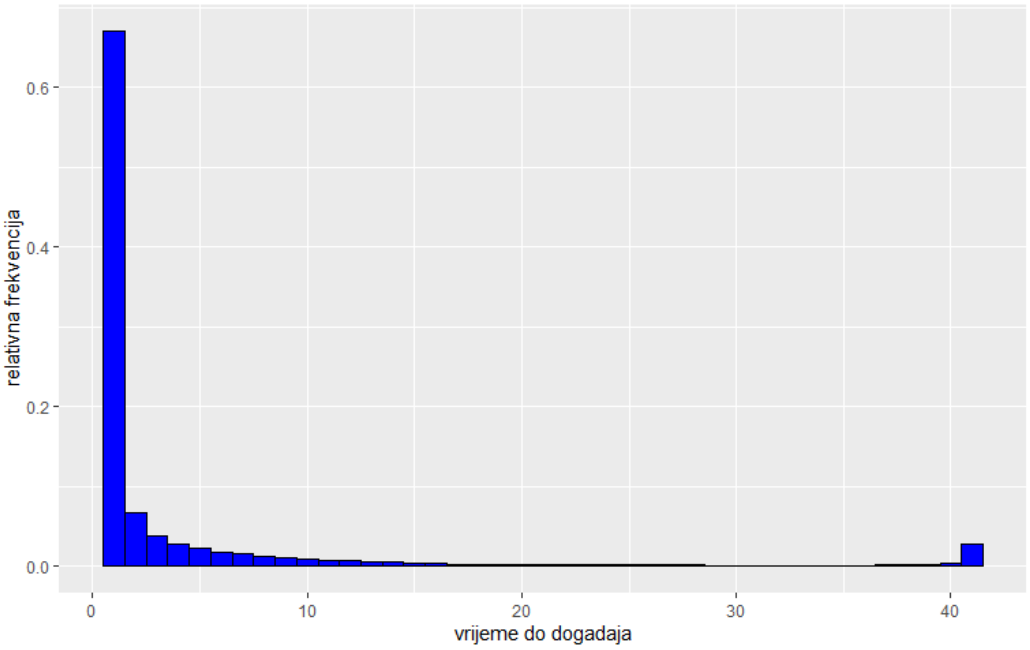
varijabla	mos	n	$\bar{x}$	s	min	q_L	medijan	q_U	max
t događaj	android	37554	4.12	8.14	1	1	1	2	41
	ios	10028	5.88	10.5	1	1	1	4	41
t kupnja	android	839	4.42	6.26	1	1	2	5	41
	ios	269	4.55	5.77	1	1	2	6	39

Tablica 2.3: Opisna statistika za varijable grupirane po kategorijama varijable *država*

varijabla	*država*	n	$\bar{x}$	s	min	q_L	medijan	q_U	max
t događaj	br	2391	1.77	3.71	1	1	1	1	41
	de	2758	5.33	9.94	1	1	1	4	41
	es	4425	5.30	9.73	1	1	1	4	41
	fr	3895	5.54	9.92	1	1	1	4	41
	gb	1994	4.64	8.79	1	1	1	3	41
	in	3718	3.75	7.69	1	1	1	2	41
	it	2814	5.26	9.89	1	1	1	3	41
	mx	2379	3.34	6.97	1	1	1	2	41
	ru	15913	4.53	8.67	1	1	1	3	41
	us	7295	4.37	8.44	1	1	1	3	41
t kupnja	br	5	4	4.24	1	1	1	7	10
	de	160	3.91	5.84	1	1	2	4	40
	es	61	5.92	7.45	1	1	3	9	37
	fr	196	3.89	6.09	1	1	1	3	39
	gb	48	3.77	5.63	1	1	1.5	3.25	30
	in	18	10.3	10.3	1	2.25	5.5	15.8	35
	it	88	4.39	4.81	1	1	2	6	24
	mx	12	6.83	6.87	1	2	4.5	8.25	24
	ru	116	6.53	6.44	1	2	5	9	39
	us	404	3.89	5.80	1	1	1	4	41

Tablica 2.4: Opisna statistika za varijable grupirane po kategorijama varijable kampanja

varijabla	*kampanja*	n	$\bar{x}$	s	min	q_L	medijan	q_U	max
t događaj	AEO	905	3.55	7.56	1	1	1	2	41
	MAI	16054	4.41	8.51	1	1	1	3	41
	VO	4163	5.80	10.8	1	1	1	4	41
	other	5461	4.77	9.15	1	1	1	3	41
	organic	20999	4.27	8.32	1	1	1	3	41
t kupnja	AEO	25	4.32	4.88	1	1	2	7	20
	MAI	87	5.91	5.73	1	2	5	7.5	39
	VO	426	3.83	5.67	1	1	1	4	41
	other	347	3.92	6.24	1	1	1	4	40
	organic	223	5.91	6.83	1	1	3	8	34



Slika 2.1: Histogram relativnih frekvencija za varijablu *t događaj*

su korisnici koji su dnevno aktivni i vjerojatno nastavljaju igrati i nakon promatranog vremena. Na slici 2.2 vidimo da je više od 40% korisnika koji su nešto kupili u promatranom vremenu prvu kupnju obavilo prvi dan igranja.

Tablica 2.5: Broj opažanja i udio cenzuriranih opažanja

varijabla		ukupno	cenzuriranja	postotak cenzuriranja
		47582	2056	4.32%
mos	android	37554	1354	3.61%
	ios	10028	702	7.00%
država	br	2391	13	0.54%
	de	2758	175	6.35%
	es	4425	248	5.60%
	fr	3895	233	5.98%
	gb	1994	88	4.41%
	in	3718	112	3.01%
	it	2814	168	5.97%
	mx	2379	58	2.44%
	ru	15913	664	4.17%
	us	7295	297	4.07%
kampanja	AEO	905	28	3.09%
	MAI	16054	649	4.04%
	organic	20999	793	3.78%
	other	5461	274	5.02%
	VO	4163	312	7.49%

Ako gledamo varijable ovisno o operacijskom sustavu (tablica 2.2), vidimo da korisnika Androida ima znatno više nego korisnika iOS-a. Aritmetička sredina vremena do događaja korisnika sustava iOS je veća. Postotak korisnika koji nisu odustali od igre je 3.61% za Android korisnike u odnosu na 7.00% za iOS korisnike. Platitelja na iOS operacijskom sustavu ima malo više, 2.68% u odnosu na 2.23% na Androidu.

Gledajući varijable po državama (tablica 2.3), vidimo da se prosječno vrijeme do događaja korisnika razlikuje i da je vidljivo najmanje u Brazilu gdje je u prosjeku 1.77 dan do odustajanja ili cenzuriranja i samo 0.54% korisnika nije odustalo od igre. Države s najvećim prosječnim vremenom do odustajanja su Njemačka, Španjolska, Francuska i Italija gdje su i najveći postoci korisnika koji nisu odustali od igre. Najviše korisnika u uzorku je iz Rusije, čak 15913 što je trećina uzorka. Postotak korisnika koji su kupili nešto u promatranom vremenu također se razlikuje između država i najveći je u Njemačkoj, Francuskoj i Sjedinjenim Američkim Državama, a najmanji u Brazilu, Meksiku i Rusiji.

Gledajući varijable po kampanjama (tablica2.4), vidimo da je najviše organskih koris-

Tablica 2.6: Broj platitelja i udio platitelja u uzorku.

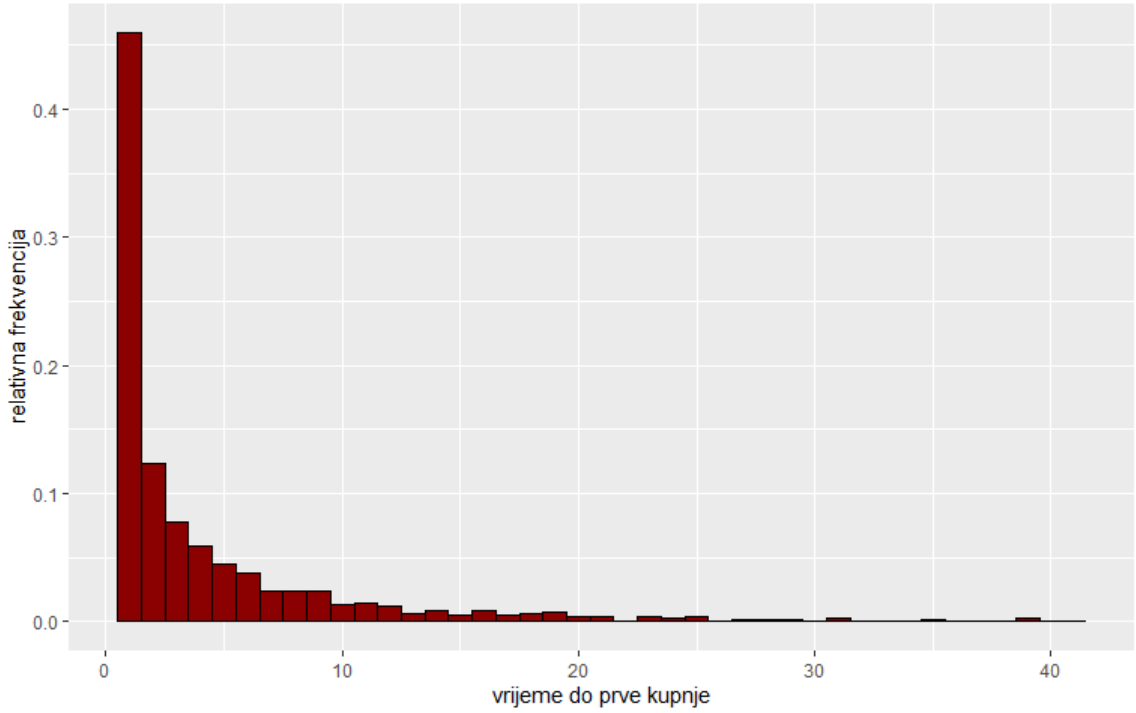
varijabla		ukupno	platitelji	postotak platitelja
		47582	1108	2.33%
mos	android	37554	839	2.23%
	ios	10028	269	2.68%
država	br	2391	5	0.21%
	de	2758	160	5.80%
	es	4425	61	1.38%
	fr	3895	196	5.03%
	gb	1994	48	2.41%
	in	3718	18	0.48%
	it	2814	88	3.13%
	mx	2379	12	0.50%
	ru	15913	116	0.73%
	us	7295	404	5.54%
kampanja	AEO	905	25	2.76%
	MAI	16054	87	0.54%
	organic	20999	223	1.06%
	other	5461	347	6.35%
	VO	4163	426	10.23%

nika koji nisu iz kampanje pa onda iz MAI kampanja. Najveće prosječno vrijeme korisnika je iz VO kampanja, a najmanje iz AEO kampanja. Najveći postotak korisnika koji nisu odustali od igre je iz VO kampanja gdje je i najveći postotak platitelja, dok korisnici iz MAI kampanja imaju najmanji postotak platitelja (vidi tablicu 2.6).

Uzorak se sastoji od 47582 korisnika iz deset različitih država, dva operacijska sustava i pet različitih vrsta kampanja. U tablici 2.7 vidimo da su u Brazilu korisnici većinom organski na operacijskom sustavu Android. Najviše je korisnika iz Rusije i oni su većinom iz MAI kampanje ili organski. Organskih korisnika je puno više na operacijskom sustavu Android nego na operacijskom sustavu iOS. VO kampanja je zastupljena u svim državama, ali najmanje u Brazilu i Rusiji. Korisnici iz AEO kampanja su većinom iz Francuske, Velike Britanije i SAD-a.

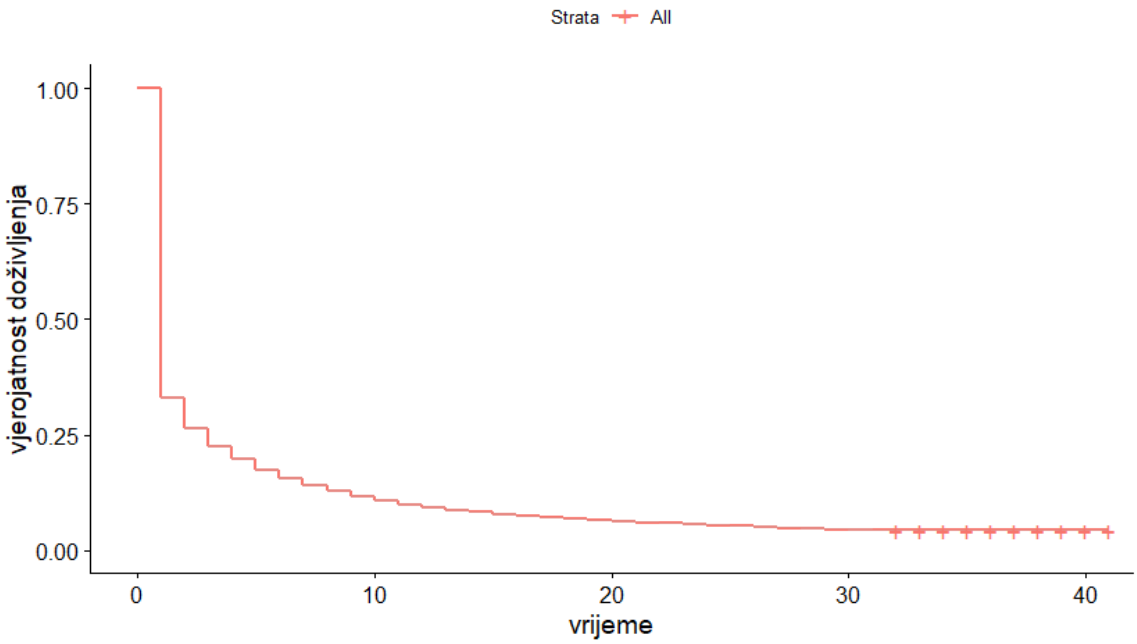
Tablica 2.7: Tablica frekvencija po kategorijskim varijablama *mos*, *država*, *kampanja*

		AEO	MAI	organic	other	VO
br	android	0	0	2290	1	1
	ios	0	0	33	21	45
de	android	5	292	828	988	262
	ios	4	108	125	18	128
es	android	1	2147	1328	30	322
	ios	4	141	188	3	261
fr	android	140	267	1082	1240	381
	ios	149	182	215	35	204
gb	android	30	1	416	16	114
	ios	221	740	297	36	123
in	android	47	0	2974	13	491
	ios	81	0	54	10	48
it	android	1	1036	899	24	344
	ios	4	206	148	14	138
mx	android	2	1	2024	14	273
	ios	5	1	26	7	26
ru	android	0	7011	4562	20	2
	ios	6	3626	612	33	41
us	android	17	0	2136	2827	654
	ios	188	295	762	111	305



Slika 2.2: Histogram relativnih frekvencija za varijablu *t kupnja*

Usporedba Kaplan-Meierovih procjena funkcija doživljenja

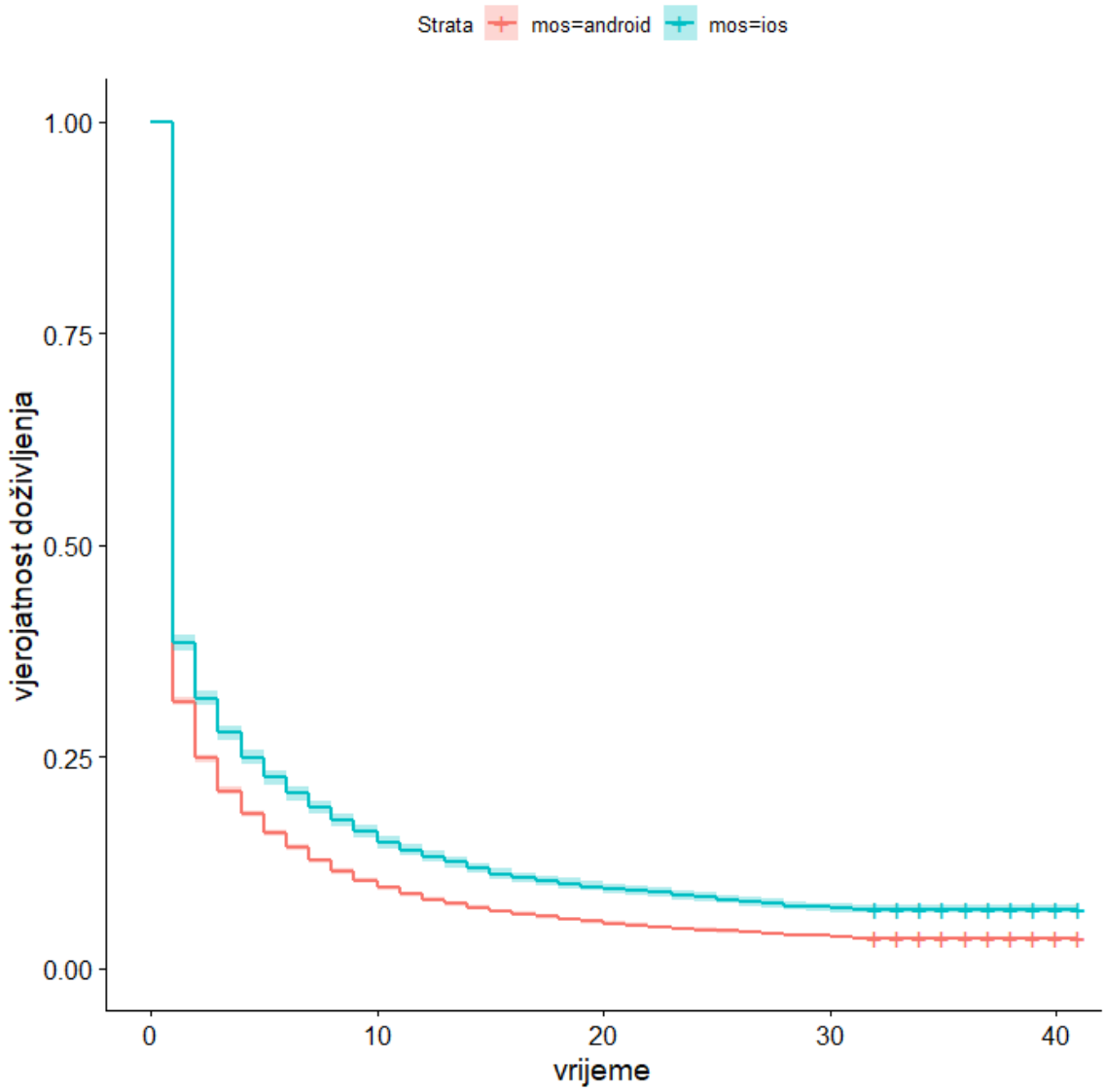


Slika 2.3: Kaplan-Meierova procjena funkcije doživljenja za cijeli uzorak

Kaplan-Meierova procjena funkcije doživljenja (Slika 2.3) pokazuje najveći skok u prvom danu gdje čak 66% korisnika ima zabilježen događaj. Cenzurirana opažanja su od 32. do 41. dana igranja. U 30. danu je 5% korisnika pod rizikom, tj. još nije odustalo od igre.

Iz procijenjenih funkcija doživljenja po operacijskim sustavima (Slika 2.4) vidimo da je vjerojatnost doživljenja veća za korisnike na iOS-u u svim danima. Gledajući po državama (Slika 2.5), korisnici iz Brazila imaju vidno najmanju vjerojatnost doživljenja u svim danima igranja, dok se za ostale države ne vidi jasna razlika. Gledajući po kampanjama (Slika 2.6), korisnici iz VO kampanja imaju najveću vjerojatnost doživljenja u svim danima igranja. U tablici 2.8 su rezultati *log-rank* testa za sve kategorijske varijable. Za sve tri varijable odbacujemo hipotezu o jednakosti funkcija doživljenja na razini značajnosti od 5%.

Daljnjim testiranjem (tablica 2.9) utvrđeno je da se doživljenje u Brazilu statistički značajno razlikuje od doživljenja u svim drugim državama. Doživljenje u Indiji se statistički značajno razlikuje od doživljenja u svim drugim državama osim Meksika i Meksiko se razlikuje od svih osim Indije. Doživljenje u SAD-u, Velikoj Britaniji i Rusiji međusobno



Slika 2.4: Kaplan-Meierova procjena funkcija doživljenja po kategorijama varijable *mos*

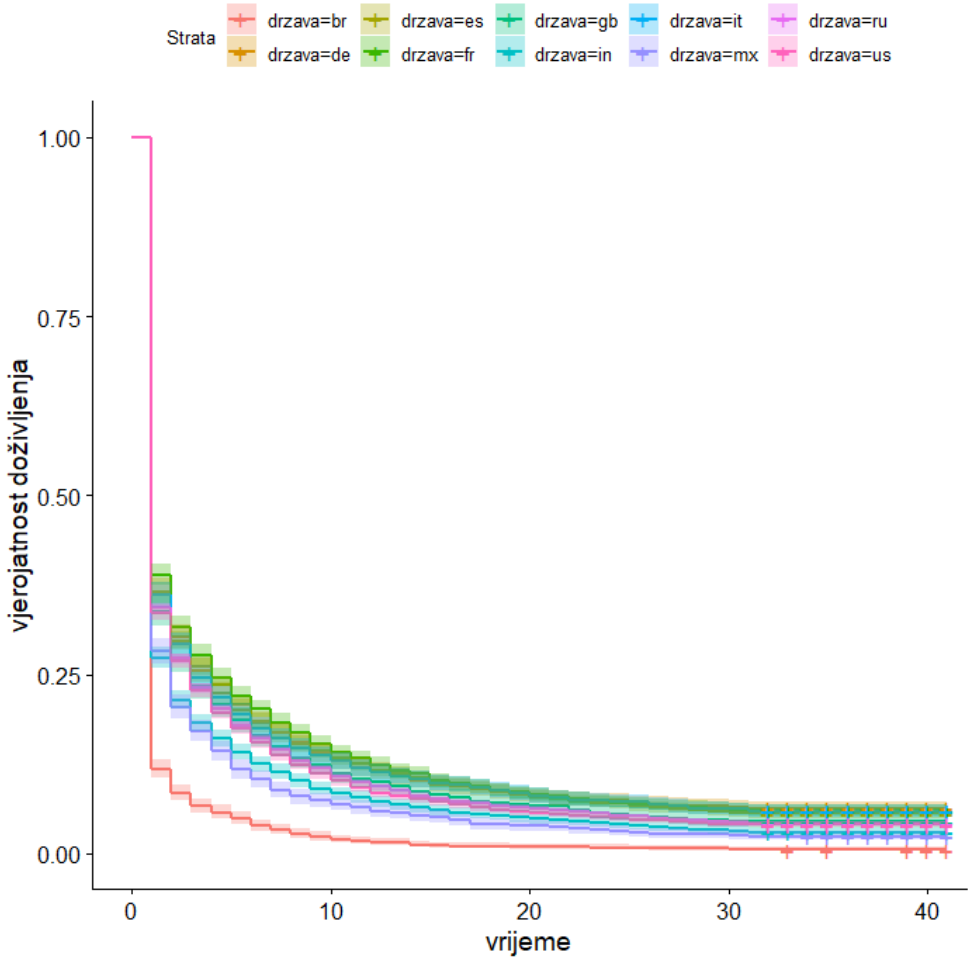
se ne razlikuje statistički značajno. Doživljenje korisnika iz Njemačke, Španjolske, Francuske i Italije međusobno se ne razlikuje statistički značajno i to su države s najvećom vjerojatnošću doživljenja korisnika u promatranom vremenu. Doživljenje korisnika iz Velike Britanije ne razlikuje se statistički značajno od doživljenja korisnika iz Rusije, SAD-a, Njemačke, Španjolske i Italije, ali se razlikuje od doživljenja korisnika iz Francuske. Francuska ima najveću prosječnu vrijednost vremena doživljenja (tablica 2.3). No,

Tablica 2.8: Rezultati *log-rank* testa za usporedbu funkcija doživljenja po kategorijama promatranih varijabli.

varijabla	χ^2	df	p-value
mos	305	1	<0.001
država	757	9	<0.001
kampanja	110	4	<0.001

uspoređujući postotke korisnika koji nisu odustali od igre u promatranom vremenu (tablica 2.5), najveći postotak je u Njemačkoj, a zatim u Francuskoj pa je vjerojatnost doživljenja na kraju promatranog vremena veća za korisnike iz Njemačke.

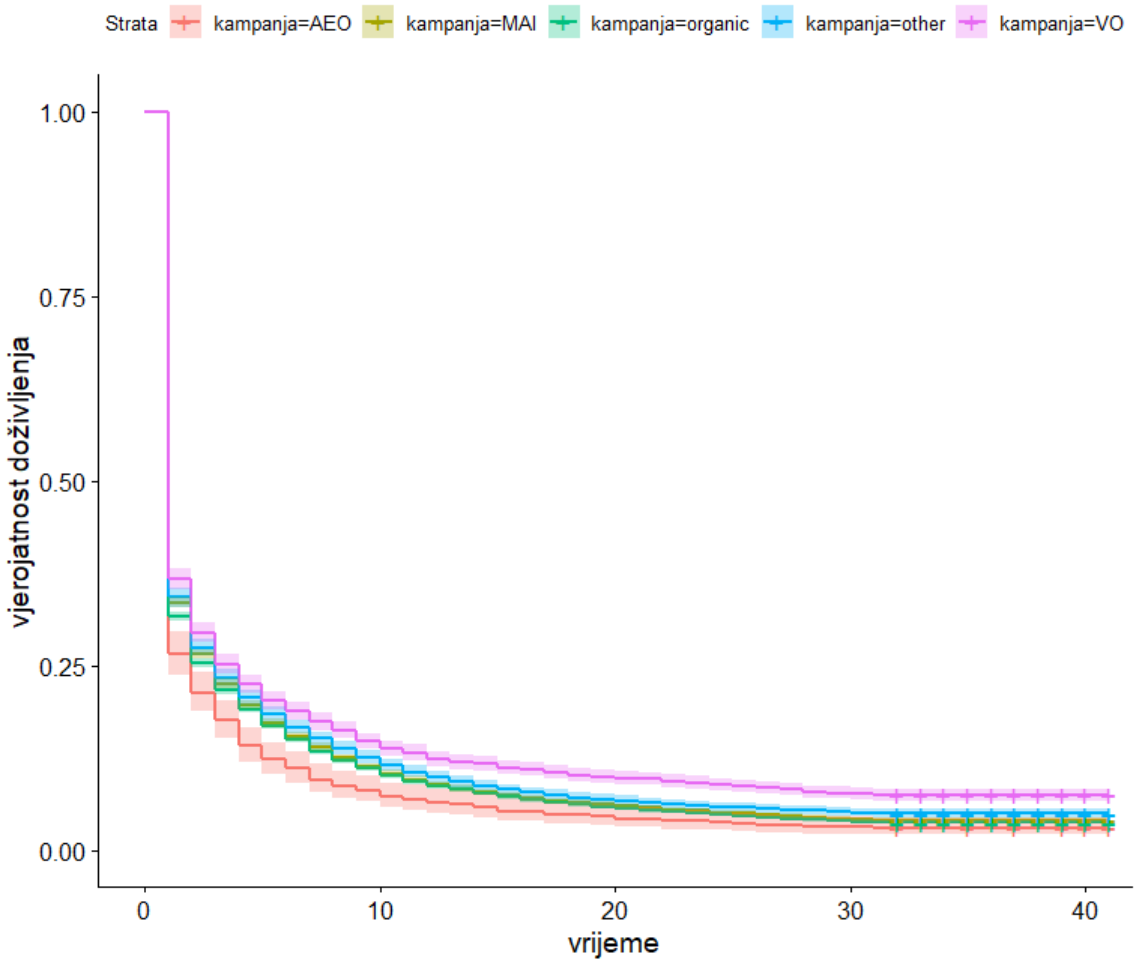
Doživljenje korisnika iz VO kampanja statistički se značajno razlikuje od doživljenja svih ostalih grupa i procijenjena vjerojatnost doživljenja je najveća za njih. Doživljenje organskih korisnika statistički se značajno razlikuje od doživljenja korisnika iz AEO kampanja, VO kampanja i ostalih (other) kampanja, a ne razlikuje se od korisnika iz MAI kampanja. Doživljenje korisnika iz MAI kampanja ne razlikuje se statistički značajno od doživljenja organskih korisnika i korisnika iz ostalih (other) kampanja. Doživljenje korisnika iz AEO kampanja statistički se značajno razlikuje od doživljenja svih ostalih grupa i procijenjena vjerojatnost doživljenja je najmanja za njih u svim danima. Rezultati testova su u tablici 2.10.



Slika 2.5: Kaplan-Meierova procjena funkcija doživljenja po kategorijama varijable *država*

Tablica 2.9: *p*-vrijednosti testa s Bonferronijevom korekcijom za varijablu *država*

	br	de	es	fr	gb	in	it	mx	ru
de	<0.001	-	-	-	-	-	-	-	-
es	<0.001	1	-	-	-	-	-	-	-
fr	<0.001	1	1	-	-	-	-	-	-
gb	<0.001	0.4194	0.41457	0.00792	-	-	-	-	-
in	<0.001	<0.001	<0.001	<0.001	<0.001	-	-	-	-
it	<0.001	1	1	1	1	<0.001	-	-	-
mx	<0.001	<0.001	<0.001	<0.001	<0.001	1	<0.001	-	-
ru	<0.001	<0.001	<0.001	<0.001	1	<0.001	0.0113	<0.001	-
us	<0.001	<0.001	<0.001	<0.001	1	<0.001	0.0021	<0.001	1



Slika 2.6: Kaplan-Meierova procjena funkcija doživljenja po kategorijama varijable *kampanja*

Tablica 2.10: *p*-vrijednosti *log-rank* testa s Bonferronijevom korekcijom za varijablu kampanja

	AEO	MAI	organic	other
MAI	0.00174	-	-	-
organic	0.01765	0.47898	-	-
other	<0.001	0.107	0.00066	-
VO	<0.001	<0.001	<0.001	<0.001

Coxov regresijski model

Iz procjenjenih funkcija doživljenja i *log-rank* testa opaženo je da korisnici na operacijskom sustavu iOS imaju veću vjerojatnost doživljenja, a pomoću univarijatnog Coxovog regresijskog modela procijenjen je efekt varijable *mos* koji je statistički značajno različit od 0 (tablica 2.11). Korisnici operacijskog sustava iOS imaju 17.7% manji hazard odustajanja od igre od korisnika Androida. Da bi model vrijedio, trebala bi biti zadovoljena pretpostavka proporcionalnosti hazarda tj. da se omjer hazarda dva različita subjekta ne mijenja kroz vrijeme. U modelu u kojem je uvedena interakcija varijable *mos* i $\ln(t)$, koeficijent uz interakciju je statistički značajno različit od 0 ($p = 0.027$) pa je pretpostavka proporcionalnosti hazarda odbačena (tablica 2.12). Za utvrđivanje dana u kojem se mijenja omjer hazarda, procijenjeni su modeli sa svim različitim vrijednostima u kojima korisnici odustaju od igre $\tau = 1, 2, \ldots, 31$ i model s najvećom vrijednosti funkcije parcijalne vjerodostojnosti je onaj s $\tau = 1$ pa je on primijenjen. Sada su koeficijenti modela različiti u vremenima $t_1 \in \langle 0, 1]$ i $t_2 \in \langle 1, 31]$. Za dobiveni model provjerena je pretpostavka proporcionalnosti hazarda istom metodom za period t_2 i pretpostavka nije odbačena ($p = 0.627$) na razini značajnosti od 5%.

Dobiveno je da je omjer hazarda odustajanja u prvom danu igranja iOS korisnika u odnosu na Android korisnike 0.839 dok je u ostalim danima taj omjer manji i iznosi 0.791, što znači da se hazard odustajanja više razlikuje u drugom vremenskom periodu.

Tablica 2.11: Univarijatni model s varijablom *mos*, $LL = -451113.5$

	$\hat{\beta}$	st.pogreška	z	p	HR	95% pouzdan interval	
mos ios	-0.195	0.012	-16.770	<0.001	0.823	0.804	0.842

Referetna vrijednost varijable *mos* je android.

Tablica 2.12: Model za testiranje proporcionalnosti hazarda, $LL = -451111.1$

	$\hat{\beta}$	st.pogreška	z	p	HR	95% pouzdan interval	
mos ios	-0.179	0.014	-13.170	<0.001	0.836	0.814	0.858
mos ios× ln(*t događaj*)	-0.028	0.013	-2.215	0.027	0.973	0.949	0.997

Referetna vrijednost varijable *mos* je android.

Procijenjen je multivarijatni model s varijablama: *mos* čija je referentna vrijednost android, *država* čija je referentna vrijednost us jer je SAD najveće tržište mobilnih igara od država u kojima je igra dostupna, *kampanja* čija je referentna vrijednost organic koju imaju

Tablica 2.13: Vrijednosti funkcije parcijalne log-vjerodostojnosti za različite τ.

τ	parcijalna log-vjerodostojnost
1	-451110.7
2	-451111.7
3	-451111.3
4	-451111.6
5	-451112.1
6	-451112.7
7	-451113.2
8	-451113.2
9	-451113.2
10	-451112.9
11	-451113.0
12	-451113.0
13	-451112.9
14	-451112.4
15	-451111.4

Tablica 2.14: Univarijatni model s koeficijentima promjenjivim u vremenu $\tau = 1$, $LL = -451110.7$

	$\hat{\beta}$	st.pogreška	z	p	HR	95% pouzdan interval	
mos ios t_1	-0.176	0.014	-12.370	<0.001	0.839	0.816	0.863
mos ios t_2	-0.235	0.020	-11.550	<0.001	0.791	0.760	0.823

Referetna vrijednost varijable *mos* je android.

tzv. organski korisnici koji nisu instalirali igru pomoću kampanje nego npr. traženjem igre u trgovini igara i *platitelj* koja je ovisna o vremenu. Koeficijent uz varijablu *mos* se nije puno promijenio dodavanjem drugih varijabli u model pa nemamo razloga sumnjati u problem združivanja za tu varijablu. Za varijable *mos*, *država* i *kampanja* testirana je pretpostavka proporcionalnosti hazarda uvođenjem interakcije varijable s vremenom i pretpostavka je odbačena za sve tri varijable (tablica 2.14 i tablica 2.15). Za sve varijable je uzeto vrijeme $\tau = 1$ u kojem je pretpostavljeno da se omjer hazarda mijenja i procijenjen je model s istim varijablama i koeficijentima promjenjivim u vremenu $\tau = 1$.

Omjer hazarda najveći je između Brazila i SAD-a što znači da kao što je već viđeno, korisnici iz Brazila imaju veći hazard odustajanja od korisnika iz drugih država. Hazard je 66.9% veći za korisnike iz Brazila u odnosu na korisnike iz SAD-a u prvom danu igranja. Razlog tome bi moglo biti to što u promatranom vremenu igra nije bila dostupna na

Tablica 2.15: Rezultati testa omjera vjerodostojnosti za usporedbu modela sa i bez interakcije s vremenom

	χ^2	df	p
država	21.354	9	0.01117
kampanja	31.561	4	<0.001

portugalskom jeziku, dok je za druge promatrane države igra bila dostupna na službenom jeziku države. Efekt je manji u periodu od drugog dana nadalje, hazard je tada 24.8% veći za korisnike u Brazilu u odnosu na korisnike u SAD-u. Hazard je veći za korisnike iz Meksika prvi dan igranja, ali u ostalim danima ne se razlikuje statistički značajno od hazarda odustajanja korisnika iz SAD-a. Hazard odustajanja za korisnike iz Indije je 14.5% veći prvi dan igranja, a ostale dane 14.9% manji od hazarda korisnika iz SAD-a. Iz toga vidimo da korisnici iz Indije koji ostanu u igri nakon prvog dana dulje igraju od korisnika iz SAD-a. Hazard odustajanja korisnika iz Njemačke, Španjolske, Francuske, Italije i Rusije je statistički značajno manji od hazarda korisnika iz SAD-a. Hazard korisnika iz Velike Britanije se ne razlikuje statistički značajno od hazarda korisnika iz SAD-a. To nam ukazuje na slično ponašanje u igri korisnika iz tih dviju država. U usporedbi s organskim korisnicima, korisnici iz AEO kampanja imaju veći hazard odustajanja za 36.3% prvi dan i 33.1% u ostalim danima. Korisnici iz MAI kampanja imaju 12% veći hazard u svim danima u odnosu na organske korisnike. Procijenjeni koeficijent ostalih kampanja i VO kampanja nije statistički značajno različit od 0 tj. korisnici iz tih kampanja nemaju statistički značajno različit hazard od organskih korisnika. Rezultati su drugačiji u odnosu na tablicu 2.10 jer je efekt varijable *kampanja* djelomično objašnjen efektom ostalih varijabli, npr. ako je korisnik iz VO kampanje kojoj je cilj prikupljati platitelje, hazard koji se razlikovao između organskih korisnika i korisnika iz VO kampanja sada je objašnjen varijablom *platitelj*. Hazard odustajanja korisnika koji su platitelji iznosi 0.402 hazarda odustajanja korisnika koji nisu platitelji. Korisnici koji nisu platitelji imaju 2.49 puta veći hazard odustajanja.

Za sve navedene usporedbe između subjekata s različitim vrijednostima promatrane varijable, pretpostavljeno je da su vrijednosti ostalih varijabli ostale nepromijenjene.

U model s varijablama *platitelj*, *mos*, *kampanja* i *država* uvedene su interakcije koje statistički značajno pospješuju model, a to su interakcije između varijabli *mos* i *država*, *mos* i *kampanja*, *kampanja* i *država*. Tako je dobiven finalni model. U tablici 2.18 prikazani su procijenjeni koeficijenti uz interakcije samo onih kategorija varijabli za koje su koeficijenti statistički značajno različiti od 0.

U finalnom su modelu procijenjeni koeficijenti uz neke kategorije varijabli promijenjeni u odnosu na prošli model jer je dio efekta objašnjen interakcijama između varijabli. U finalnom je modelu procijenjeni koeficijent uz kategoriju ios varijable *mos* jednak -0.471 tj. hazard odustajanja korisnika iOS-a je 37.5% manji u odnosu na korisnike An-

Tablica 2.16: Multivarijatni model s varijablama *mos*, *država*, *kampanja*, *platitelj*, $LL = -450472.3$

	$\hat{\beta}$	st.pogreška	z	p	HR	95% p.i.	
mos ios	-0.214	0.013	-16.834	<0.001	0.807	0.787	0.828
država br	0.451	0.025	17.848	<0.001	1.569	1.494	1.649
država de	-0.128	0.023	-5.549	<0.001	0.880	0.841	0.920
država es	-0.187	0.022	-8.633	<0.001	0.829	0.795	0.865
država fr	-0.173	0.021	-8.422	<0.001	0.841	0.808	0.876
država gb	-0.003	0.027	-0.106	0.916	0.997	0.945	1.052
država in	0.052	0.022	2.363	0.018	1.053	1.009	1.099
država it	-0.147	0.024	-6.034	<0.001	0.863	0.823	0.906
država mx	0.089	0.025	3.543	<0.001	1.093	1.041	1.149
država ru	-0.101	0.018	-5.637	<0.001	0.904	0.873	0.936
kampanja AEO	0.304	0.036	8.463	<0.001	1.356	1.263	1.455
kampanja MAI	0.112	0.013	8.579	<0.001	1.119	1.090	1.148
kampanja other	0.010	0.018	0.537	0.591	1.010	0.975	1.046
kampanja VO	-0.018	0.018	-1.001	0.317	0.982	0.948	1.018
platitelj	-0.878	0.048	-18.269	<0.001	0.416	0.378	0.457

Referetne vrijednosi varijabli: *mos* = android, *država* = us, *kampanja* = organic.

droida ako varijable *kampanja* i *država* postižu referentnu vrijednost, a varijabla *platitelj* ostaje nepromijenjena. Za korisnike iz Velike Britanije, efekt se mijenja i koeficijent iznosi $-0.471 + 0.219 = -0.252$ tj. korisnici iOS-a iz Velike Britanije imaju 22.3% manji hazard odustajanja od korisnika Androida iz Velike Britanije uz referentnu vrijednost varijable *kampanja* i nepromijenjenu vrijednost varijable *platitelj*. Za korisnike iOS-a iz Velike Britanije i VO kampanje koeficijent iznosi $-0.471 + 0.219 + 0.391 = 0.139$ tj. oni imaju 14.9% veći hazard odustajanja od korisnika Androida iz Velike Britanije i VO kampanje uz varijablu *platitelj* nepromijenjenu. Korisnici iz Rusije u odnosu na korisnike iz SAD-a nemaju statistički značajno različit hazard odustajanja ako varijable *mos* i *kampanja* postižu referentnu vrijednost, a varijabla *platitelj* ostaje nepromijenjena. No, korisnici iz Rusije iz MAI kampanje imaju statistički značajno manji hazard od korisnika iz SAD-a iz MAI kampanje, a to vrijedi i za korisnike ostalih (other) kampanja, korisnike VO kampanja, kao i za korisnike iOS-a iz Rusije u odnosu na korisnike iOS-a iz SAD-a u drugom vremenskom intervalu. Dakle, hazard odustajanja za korisnike iz Rusije i SAD-a se razlikuje samo za dio korisnika u drugom vremenskom intervalu. Korisnici iz Velike Britanije sada imaju statističko značajno različit hazard od korisnika iz SAD-a dok ostale varijable postižu referentnu vrijednost i on je za 13.7% manji prvi dan i 21.8% manji u ostalim danima igranja.

Tablica 2.17: Multivarijatni model varijablama mos, *država*, *kampanja*, *platitelj* s koeficijentima promjenjivim u vremenu $\tau = 1$, $LL = -450439.4$

	$\hat{\beta}$	st.pogreška	z	p	HR	95% p.i.	
mos ios t_1	-0.177	0.016	-11.330	<0.001	0.838	0.812	0.864
mos ios t_2	-0.287	0.022	-13.158	<0.001	0.750	0.719	0.783
država br t_1	0.512	0.028	18.204	<0.001	1.669	1.580	1.764
država br t_2	0.222	0.066	3.347	<0.001	1.248	1.096	1.421
država de t_1	-0.095	0.028	-3.392	<0.001	0.909	0.861	0.961
država de t_2	-0.208	0.041	-5.074	<0.001	0.812	0.749	0.880
država es t_1	-0.123	0.026	-4.648	<0.001	0.885	0.840	0.932
država es t_2	-0.327	0.038	-8.576	<0.001	0.721	0.669	0.777
država fr t_1	-0.164	0.025	-6.503	<0.001	0.849	0.808	0.892
država fr t_2	-0.207	0.036	-5.816	<0.001	0.813	0.758	0.872
država gb t_1	0.023	0.033	0.691	0.489	1.023	0.959	1.091
država gb t_2	-0.070	0.048	-1.448	0.148	0.932	0.848	1.025
država in t_1	0.136	0.026	5.264	<0.001	1.145	1.089	1.205
država in t_2	-0.161	0.042	-3.862	<0.001	0.851	0.784	0.924
država it t_1	-0.096	0.030	-3.242	0.001	0.909	0.858	0.963
država it t_2	-0.262	0.043	-6.087	<0.0012	0.770	0.708	0.837
država mx t_1	0.119	0.030	3.982	<0.001	1.126	1.062	1.194
država mx t_2	0.030	0.048	0.627	0.530	1.030	0.938	1.131
država ru t_1	-0.064	0.022	-2.939	0.003	0.938	0.899	0.979
država ru t_2	-0.183	0.032	-5.743	<0.001	0.833	0.783	0.887
kampanja AEO t_1	0.310	0.042	7.444	<0.001	1.363	1.256	1.479
kampanja AEO t_2	0.286	0.072	3.975	<0.001	1.331	1.156	1.533
kampanja MAI t_1	0.109	0.016	6.909	<0.001	1.116	1.081	1.151
kampanja MAI t_2	0.114	0.023	4.908	<0.001	1.121	1.071	1.173
kampanja other t_1	0.028	0.022	1.289	0.197	1.029	0.985	1.074
kampanja other t_2	-0.022	0.032	-0.676	0.499	0.979	0.919	1.042
kampanja VO t_1	-0.021	0.022	-0.946	0.344	0.979	0.938	1.023
kampanja VO t_2	-0.005	0.033	-0.144	0.886	0.995	0.933	1.061
platitelj	-0.912	0.049	-18.535	<0.001	0.402	0.365	0.442

Referetne vrijednosi varijabli: *mos* = android, *država* = us, *kampanja* = organic.

No, korisnici iOS-a iz Velike Britanije imaju 7.4% veći hazard prvi dan i samo 1.4% veći hazard ostale dane, što znači da su korisnici iOS-a slični u Velikoj Britaniji i SAD-u, dok korisnici Androida u Velikoj Britaniji imaju manji hazard u odnosu na korisnike u SAD-

u. Korisnici iz Njemačke, Španjolske, Francuske i Italije imaju manji hazard odustajanja od korisnika iz SAD-a dok ostale varijable postižu referentnu vrijednost. Također, razlika u hazardu između korisnika iz Francuske i SAD-a je manja za korisnike iOS-a nego za korisnike Androida u prvom danu igranja, a organski korisnici iOS-a iz Italije imaju veći hazard odustajanja od korisnika iOS-a iz SAD-a prvi dan igranja. Korisnici iz AEO kampanja imaju 49.0% veći hazard na Androidu i 87.0% na iOS-u, a korisnici iz MAI kampanja imaju 39.4% veći hazard na Androidu i 84.2% na iOS-u od organskih korisnika prvi dan igranja dok ostale varijable imaju referentnu vrijednost. Hazard korisnika ostalih vrsta kampanja ne razlikuje se statistički značajno od organskih korisnika Androida iz SAD-a, dok korisnici iOS-a iz SAD-a iz VO kampanja imaju statistički značajno veći hazard odustajanja od organskih korisnika iOS-a iz SAD-a. Korisnici iz VO kampanja iz Brazila, Španjolske, Indije, Italije, Meksika i Rusije imaju statistički značajno manji hazard od organskih korisnika iz tih zemalja, dok varijabla *mos* postiže referentnu vrijednost, a varijabla *platitelj* ostaje nepromijenjena. Korisnici koji nisu platitelji u odnosu na one koji jesu imaju 2.55 puta veći hazard dok su vrijednosti ostalih varijabli nepromijenjene.

Tablica 2.18: Finalni model koji sadrži varijable *platitelj*, *mos*, *država*, *kampanja* i interakcije *mos* × *država*, *mos* × *kampanja*, *država* × *kampanja* i koeficijente promjenjive u vremenu, $LL = -450248.7$

	$\hat{\beta}$	st.pogreška	z	p	HR	95% p.i.	
platitelj	-0.936	0.050	-18.829	<0.001	0.392	0.356	0.432
mos ios t_1	-0.471	0.049	-9.576	<0.001	0.625	0.567	0.688
mos ios t_2	-0.392	0.062	-6.337	<0.001	0.676	0.599	0.763
država br t_1	0.496	0.034	14.647	<0.001	1.643	1.537	1.756
država br t_2	0.288	0.074	3.900	<0.001	1.334	1.154	1.542
država de t_1	-0.168	0.050	-3.383	<0.001	0.845	0.767	0.932
država de t_2	-0.157	0.068	-2.302	0.021	0.854	0.747	0.977
država es t_1	-0.115	0.042	-2.737	0.006	0.892	0.821	0.968
država es t_2	-0.200	0.060	-3.331	<0.001	0.819	0.728	0.921
država fr t_1	-0.196	0.045	-4.364	<0.001	0.822	0.752	0.897
država fr t_2	-0.226	0.062	-3.667	<0.001	0.798	0.707	0.900
država gb t_1	-0.148	0.063	-2.342	0.019	0.863	0.762	0.976
država gb t_2	-0.246	0.089	-2.771	0.006	0.782	0.657	0.930
država in t_1	0.127	0.033	3.811	<0.001	1.135	1.064	1.212
država in t_2	-0.091	0.053	-1.694	0.090	0.913	0.823	1.014
država it t_1	-0.125	0.048	-2.628	0.009	0.882	0.804	0.969
država it t_2	-0.211	0.068	-3.101	0.002	0.810	0.709	0.925
država mx t_1	0.152	0.036	4.192	<0.001	1.164	1.084	1.250
država mx t_2	0.167	0.059	2.810	0.005	1.181	1.052	1.327
država ru t_1	0.008	0.031	0.268	0.789	1.008	0.949	1.071
država ru t_2	-0.062	0.047	-1.314	0.189	0.940	0.857	1.031
kampanja AEO t_1	0.399	0.128	3.126	0.002	1.490	1.160	1.914
kampanja AEO t_2	0.252	0.230	1.095	0.274	1.287	0.819	2.021
kampanja MAI t_1	0.332	0.093	3.586	<0.001	1.394	1.163	1.672
kampanja MAI t_2	0.200	0.141	1.418	0.156	1.221	0.926	1.610
kampanja other t_1	0.037	0.034	1.080	0.280	1.037	0.971	1.108
kampanja other t_2	0.051	0.052	0.979	0.327	1.053	0.950	1.167
kampanja VO t_1	-0.028	0.049	-0.581	0.561	0.972	0.883	1.070
kampanja VO t_2	0.123	0.073	1.690	0.091	1.130	0.981	1.303
mos ios × *država* fr t_1	0.174	0.075	2.318	0.020	1.190	1.027	1.378
mos ios × *država* gb t_1	0.219	0.091	2.416	0.016	1.245	1.042	1.488
mos ios × *država* gb t_2	0.260	0.120	2.173	0.030	1.297	1.026	1.640
mos ios × *država* it t_1	0.233	0.080	2.896	0.004	1.262	1.078	1.477
mos ios × *država* ru t_2	-0.179	0.084	-2.125	0.034	0.836	0.709	0.986
mos ios × *kampanja* AEO t_1	0.227	0.106	2.146	0.032	1.255	1.020	1.545
mos ios × *kampanja* MAI t_1	0.279	0.048	5.752	<0.001	1.321	1.202	1.453
mos ios × *kampanja* MAI t_2	0.328	0.065	5.061	<0.001	1.388	1.222	1.575
mos ios × *kampanja* VO t_1	0.391	0.057	6.914	<0.001	1.478	1.323	1.652
mos ios × *kampanja* VO t_2	0.316	0.078	4.073	<0.001	1.372	1.178	1.597
država it × *kampanja* AEO t_2	2.032	1.028	1.977	0.048	7.628	1.018	57.188
država es × *kampanja* MAI t_1	-0.251	0.100	-2.500	0.012	0.778	0.640	0.947
država ru × *kampanja* MAI t_2	-0.360	0.092	-3.895	<0.001	0.698	0.582	0.836
država es × *kampanja* other t_1	-0.759	0.360	-2.109	0.035	0.468	0.231	0.948
država mx × *kampanja* other t_2	-0.929	0.456	-2.036	0.042	0.395	0.162	0.966
država ru × *kampanja* other t_2	-0.586	0.273	-2.148	0.032	0.557	0.326	0.950
država br × *kampanja* VO t_2	-1.476	0.487	-3.028	0.002	0.229	0.088	0.594
država es × *kampanja* VO t_2	-0.375	0.114	-3.293	<0.001	0.687	0.550	0.859
država in × *kampanja* VO t_2	-0.297	0.123	-2.416	0.016	0.743	0.584	0.945
država it × *kampanja* VO t_1	-0.174	0.087	-2.009	0.045	0.840	0.709	0.996
država mx × *kampanja* VO t_1	-0.335	0.093	-3.581	<0.001	0.715	0.596	0.859
država mx × *kampanja* VO t_2	-0.531	0.131	-4.050	<0.001	0.588	0.455	0.760
država ru × *kampanja* VO t_1	-0.622	0.240	-2.595	0.009	0.537	0.335	0.859

Referetne vrijednosi varijabli: *mos* = android, *država* = us, *kampanja* = organic.

Zaključak

U industriji mobilnih igara bitno je što dulje zadržati korisnike u igri i zato se bilježi koliko i u kojem vremenu korisnici odustaju od igre. Metode analize doživljenja pogodne su za takve podatke. Pomoću Kaplan-Meierove procjene funkcije doživljenja i *log-rank* testa vidi se da se funkcije doživljenja razlikuju za korisnike s različitim obilježjima: mobilni operacijski sustav koji koriste, država u kojoj igraju i kampanja pomoću koje su instalirali igru. Primjenom Coxovog regresijskog modela dobiveni su i omjeri hazarda korisnika s različitim obilježjima.

Pomoću dobivenog modela, osim što možemo vidjeti efekt pojedinih varijabli, moguće je i procijeniti funkciju doživljenja i hazarda za korisnike s određenim karakteristikama. Model bi mogao procijeniti kolika je vjerojatnost korisnika da odustane sljedeći dan pa ako je vjerojatnost velika, korisniku se može ponuditi neka specijalna ponuda da ne odustane od igre. U takav prediktivni model bilo bi dobro dodati još varijabli promjenjivih u vremenu koje su vezane uz događaje u igri kako bi predikcije bile što točnije.

Razlika u hazardu odustajanja korisnika iz različitih država mogla bi biti uzrokovana različitim kulturološkim navikama i konkurencijom na tržištu pojedinih država. Ovisno o vrsti kampanje zbog koje je korisnik instalirao igru, razlikuje se njegov profil i navike. To bi mogao biti razlog zašto se hazard razlikuje između korisnika iz različitih kampanja. Hazard odustajanja platitelja u odnosu na korisnike koji nisu platitelji očekivano je manji jer korisnici koji plate nešto u igri vjerojatno su zainteresirani za nju. Također, uočeno je i da omjer hazarda odustajanja između različitih korisnika nije jednak kroz vrijeme, što također ovisi o karakteristikama korisnika. Pomoću modela vidimo da na hazard ne utječu samo pojedine karakteristike korisnika, nego i kombinacije tih karakteristika.

Tehnologije i trendovi na tržištu mobilnih igara brzo se mijenjaju, a i sama igra se neprestano razvija pa uzorak korisnika koji su instalirali igru u jednom periodu može biti sasvim drugačiji od uzorka korisnika u nekom drugom periodu. Zbog toga je dobro analizu ponoviti u raznim interesnim vremenskim periodima i tako pratiti doživljenje korisnika.

Bibliografija

[1] O. Aalen, *Nonparametric Inference for a Family of Counting Processes*, The Annals of Statistics **6** (1978), br. 4, 701–726.

[2] Cox D.R., *Regression Models and Life Tables*, Journal of the Royal Statistic Society **B** (1972), br. 34, 187–202.

[3] E. L. Kaplan i P. Meier, *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association **53** (1958), br. 282, 457–481.

[4] J.P. Klein i M.L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data*, second., 2003.

[5] T.M. Therneau, T. Lumley, E. Atkinskon i C. Crowson, *Package 'survival'*, (2020), `https://cran.r-project.org/web/packages/survival/survival.pdf`.

[6] T. Therneau, E. Atkinskon i C. Crowson, *Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model*, (2020), `https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf`.

Sažetak

U ovom radu opisane su metode analize doživljenja koje se koriste za analizu učestalosti nekog događaja zajedno s vremenom do tog događaja. Metode su primijenjene na skup opažanja korisnika mobilne igre gdje je promatrani događaj odustajanje od igre. Promatrano je kako se ishod događaja i vrijeme do događaja razlikuju između korisnika s različitim značajkama kao što su država i mobilni operacijski sustav korisnika. Te razlike analizirane su pomoću Kaplan-Meierove procjene funkcije doživljenja i primjenom Coxovog regresijskog modela.

Summary

In this master thesis, methods of survival analysis are presented, which are used to analyse occurrence of some event along with the time to the event. These methods are applied on dataset of mobile game users, where the observed event is churn of the user. It is observed how outcome of the event and time to the event are different between users with different characteristics, like country and mobile operatiing system of the user. These differences are analysed with Kaplan-Meier estimate of survival function and Cox regression model.

Životopis

Rođena sam u Zagrebu, 28. rujna 1995. Završila sam XV. gimnaziju, matematičko-informatički smjer 2014. godine i Srednju glazbenu školu Pavla Markovca, odjel za teorijske glazbene predmete 2013. godine. Preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu upisala sam 2014. godine i završila ga 2017. kada sam upisala diplomski sveučilišni studij Matematička statistika na istom fakultetu. Zimski semestar akademske godine 2018./2019. provela sam na Erasmus+ studijskom boravku na Sveučilištu u Gentu. U listopadu 2019. godine počela sam raditi u tvrtki Nanobit kao mlađa analitičarka podataka.