**Biostatistika projekat**

Kreiranje prediktivnih modela koristeći metode multivarijacione analize za problem churner-a na muzičkoj platformi

Tamara Đorđević 2018-0187

Beograd 2022

Sadržaj

[1. Identifikacija problema u poslovanju 3](#_Toc106622892)

[2. Izvori prikupljenih podataka 3](#_Toc106622893)

[2.1 Istraživački cilj 3](#_Toc106622894)

[2.2 Istraživačka pitanja 3](#_Toc106622895)

[3. Analiza podataka 4](#_Toc106622896)

[3. Drvo odlučivanja 5](#_Toc106622897)

[4. Linearna regresija 7](#_Toc106622898)

[5. Analiza preživljavanja 9](#_Toc106622899)

[6. Zaključak 12](#_Toc106622900)

[7. Reference 12](#_Toc106622901)

# 1. Identifikacija problema u poslovanju

Problem sa kojim se kompanija koja je izgradila datu muzičku platformu susreće jeste kako predvideti da li će postojeći korisnik obonoviti članstvo ili ne. U ovom slučaju koristimo termin *churn*, odnosno prestanak korišćenja date muzičke platforme i prekidanje članstva.

# 2. Izvori prikupljenih podataka

Churn analiza je posebno važno u proračunima poslovanja u sektorima kao što su osiguranje, telekomunikacije ili bankarstvo, gde je model prihoda zasnovan na pretplati. Prema istraživačima, pridobijanje novih kupaca u današnjim konkurentskim uslovima je i do 10 puta skuplje od zadržavanja postojećih kupaca [1]. To je metoda analize koja se koristi u oblastima kao što su određivanje profila postojećih kupaca, analiza odlaza kupaca i procena otišlih kupaca.

Naravno, danas kompanije imaju veći pristup podacima, koje sakupljaju kroz različite kanale komunikacije i tako mogu da naprave distinkciju koji klijent je *churner* a koji to nije. [1]

Za ovu analizu koriste se različiti statistički modeli poput drva odlučivanja, linearne regresije, naivnog Bajesa, cogx regresija kao i mnogi drugi. Nakon pregleda literature Ronan Duchemin i Ricardo Matheus, profesori sa Univerziteta Erasmus iz Holadnije su kategorizovali ove modele u odnosu na performanse koje daju u rešavanju problema churn-a.[2] Zaključak do kojeg su došli jeste da je Logistička regresija model koji se najviše koristi u analizi date teme, baš zbog jednostavnosti korišćenja ovog modela. Drugi najčešće korišćen model jeste drvo odlučivanja. Hvaljen je zbog svoje visoke interpretabilnost [3] i lakoća razumevanja [4] dok u isto vreme daje odlične rezultate. Ovaj model omogućava ljudima da lakše razumeju problem koji se dešava i daje mogućnost uključenja ostalih zainteresovanih strana. Iako je lako za interpretaciju, drvo odlučivanja zna da da nešto slabije rezultate u poređenju sa ostalim metodama.

## 2.1 Istraživački cilj

Cilj ovog istraživanja je pronaći koje su ključne karakteristike korisnika koji odlaze, odnosno, onih koji se zadržavaju na muzičkoj platformi. Takođe, potrebno je predložiti strategiju za dalje delovanje kompanije u odnosu na postavljeni cilj.

## 2.2 Istraživačka pitanja

1. Na koju starosnu grupu korisnika kompanija treba da se fokusira?
2. Da li je bolje usmeriti strategiju ka premium ili basic paketu pretplate?

# 3. Analiza podataka

Svi podaci koji se nalaze u datasetu su prikazani u sledećoj tabeli:

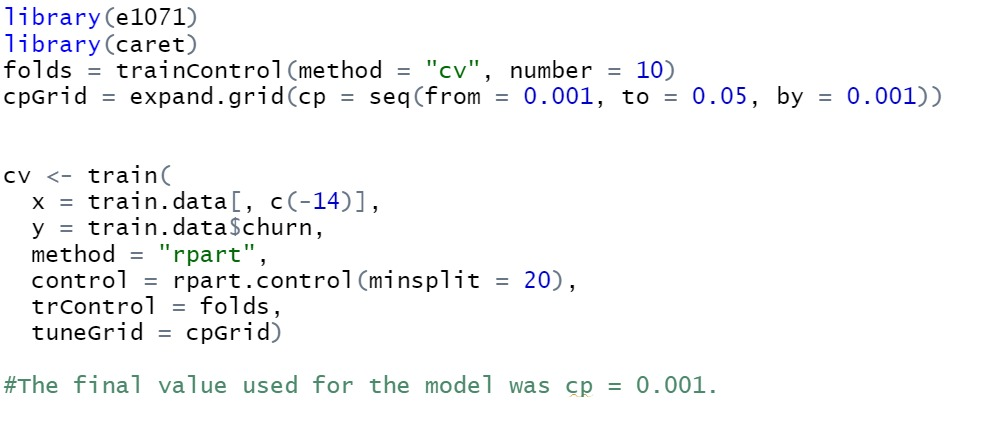
|  |  |
| --- | --- |
| Varijabla | Opis |
| id | Jedinstveni identifikator korisnika |
| datumlogovanja | Datum kada se korisnik ulogovao |
| num\_25 | Broj prekinutih pesama u prvih 25% vremena njenog trajanja |
| num\_75 | Broj prekinutih pesama tokom 50% i 75% vremena njenog trajanja |
| num\_50 | Broj prekinutih pesama u prvih 50% vremena njenog trajanja |
| num\_985 | Broj prekinutih pesama tokom 75% i 98.5% vremena njenog trajanja |
| num\_100 | Broj prekinutih pesama nakon 98.5% njenog trajanja |
| num\_unq | Broj jedinstvenih preslšanih pesama |
| ukupno\_vreme | Ukupno vreme slušanja pesama izraženo u sekundama |
| churn | Varijabla koju predviđamo |
| grad | Odakle dolaze korisnici |
| godine | Broj godina korisnika |
| pol | Pol |
| metod\_registorvanja | Metod kako su se registovali |
| datum\_registorvanja | Datum kada se korisnik registovao |
| metod\_placanja | Metoda plaćanja korisnika |
| clanstvo | Broj dana članstva |
| cena\_izabranog\_plana | Koja je cena izabranog paketa |
| uplaceno | Koliko je novca uplaćeno |
| autoproduzetak | Da li je doslšlo do automatske obnove |
| datumtransakcije | Datum kada je transakcija izvršena |
| datumisticanjaclanstva | Datum kada je isteklo članstvo |

Na osnovu datih varijabli izvedenu su:

|  |  |
| --- | --- |
| Izvedana varijabla | Opis |
| region | Da li je korisnik iz Evope ili ne |
| agegroup | Milenijalci, generacija X i Z |
| time | Broj godina od registracije |
| periodofuse | Period Koliko je koristio muzičku platformu u god |
| time\_of\_listening | Vreme slušanja muzike u minutima |
| packet\_category | * 500 – premium, u suprotnom basic |

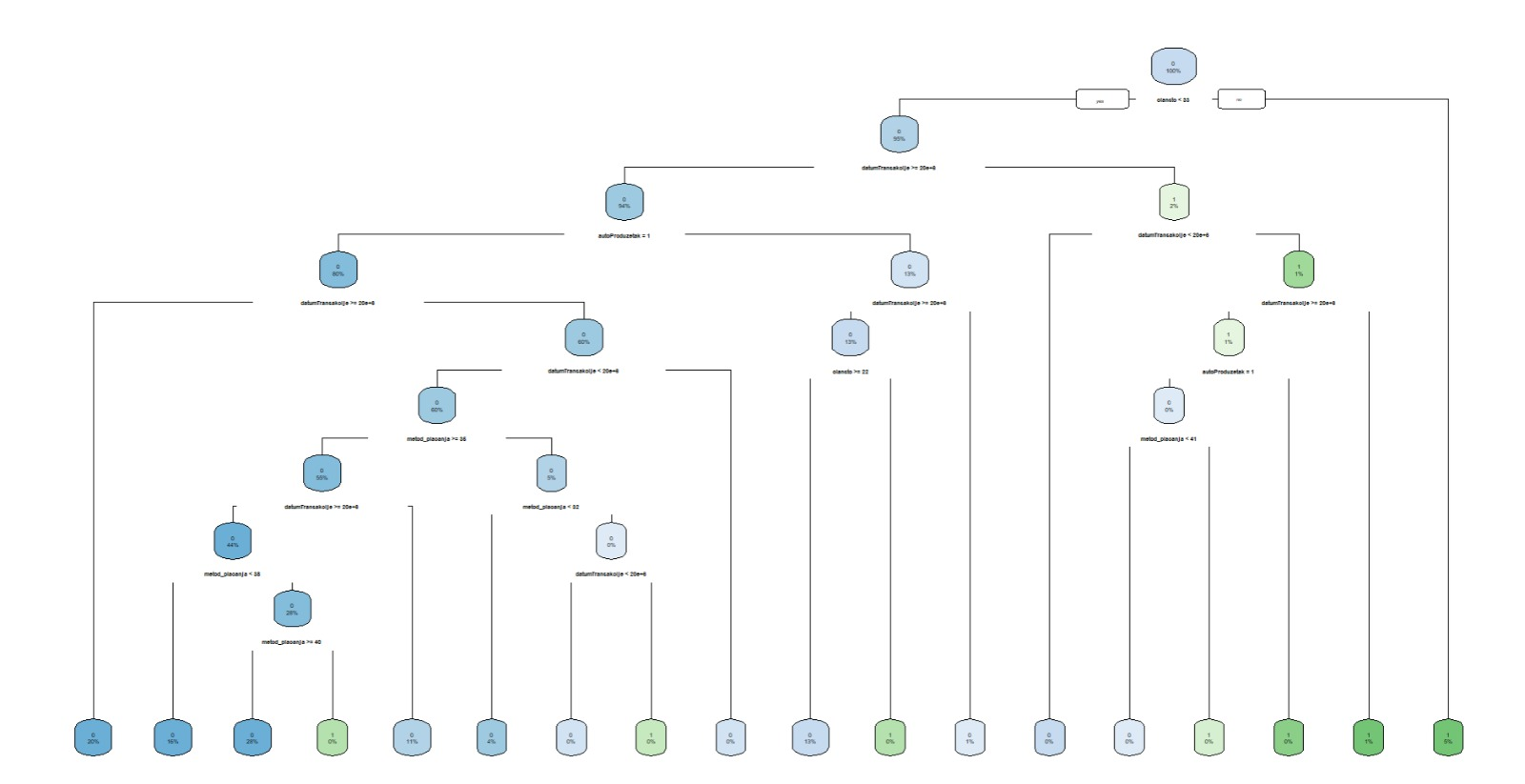
# 3. Drvo odlučivanja

Kako bismo dobili što bolje rezultate potrebno je naći odgovarajuću vrednost za cp. Iskorišćena je biblioteka caret i dobijena vrednost je 0.001.



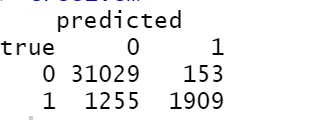
Pomoću data partition funcije napravljena su dva nova dataseta-a, train i test. Osamdeset posto podataka se nalaziu train datasetu dok ostalih 20% je u test datasetu, pri tome da imamo podjednaku raspoređenost izlazne varijable.

Drvo odlučivanja:



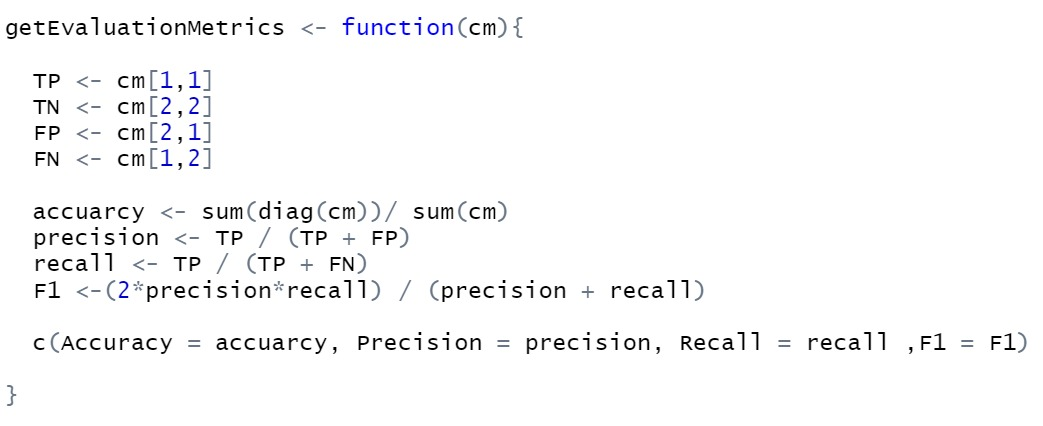
Metod plaćanja, datum transakcije, članstvo i auto produžetak su varijable koje je model prepoznao kao najbitnije.

Izračunata je matrica konfuzije:



Možemo uočiti da je naš model tačno predvideo da će 31 029 korisnika neće biti churner-i, s obzirom da je true positive 0 vrednost. Od 31 182 korisnika za 153 njih je omašio, odnosno predvideo je da će biti churer-i ali nisu ( false negative). Od ukupno 3 164 korisnika koji će napustiti platformu model je tačno predvideo za 1909 njih (true negative), a pogrešno za njih 1255.

Kako bismo odredili koliko je naš model dobar korištena je funkcija za evaluaciju metrika:



Rezultati su:



**Accuracy** je 0.959 što znači da je tačnost predviđanja našeg modela poprilično visoka. U našem datasetu to je 32 938 od ukupno 34 346.

**Precision** je udeo onih koje smo predvideli da su pozitivne i da su stvarno pozitivne i u ovom slučaju to je 0.961, odnosno 96.1%. Od ukupno 31 182 tačno predviđenih je 31 029.

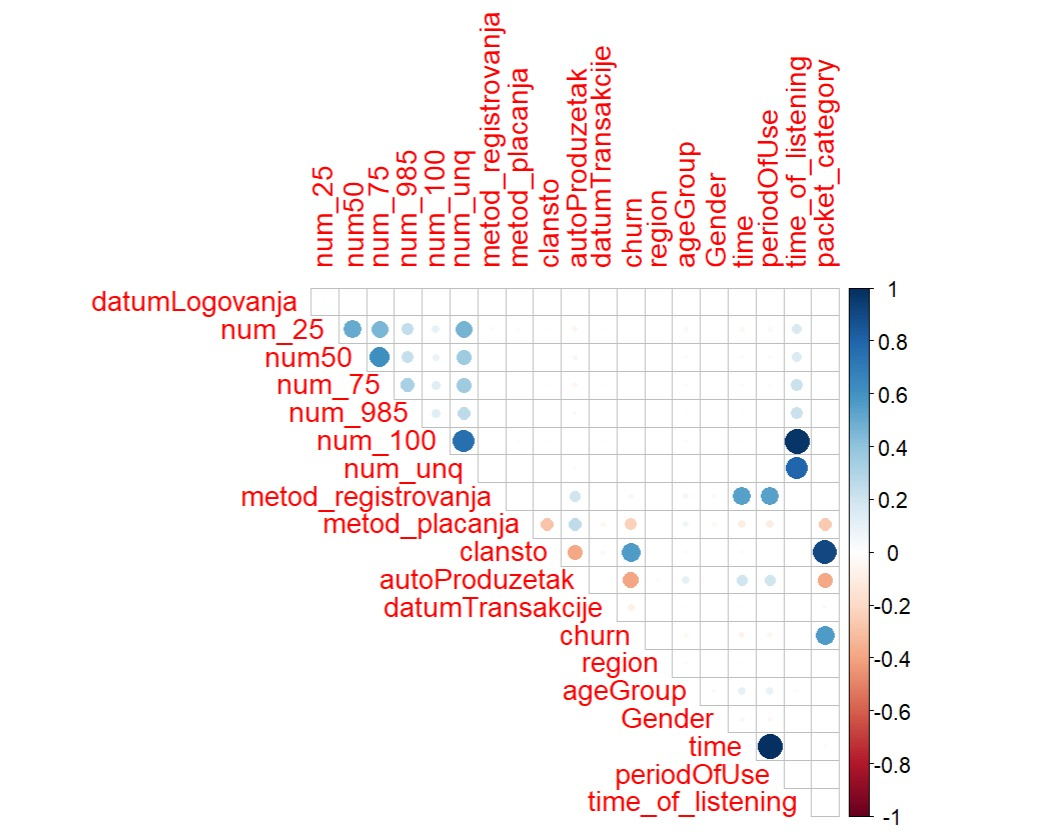
**Recall je** udeo observacija koje su stvarno stvarno pozitivne i predvideli smo da su stvarno pozitivne.Recall je jako visok i iznosi 99.5%

Na kraju **F1 statistika** govori o balansu recall i precision metrike. F1 statistika pokazuje koliko je model dobar, vrednost je 97.7% pa možemo zaključiti da je model dobar. Odnosno u 97.7% možemo odrediti da li će korisnik ostati ili će prekinuti članstvo.

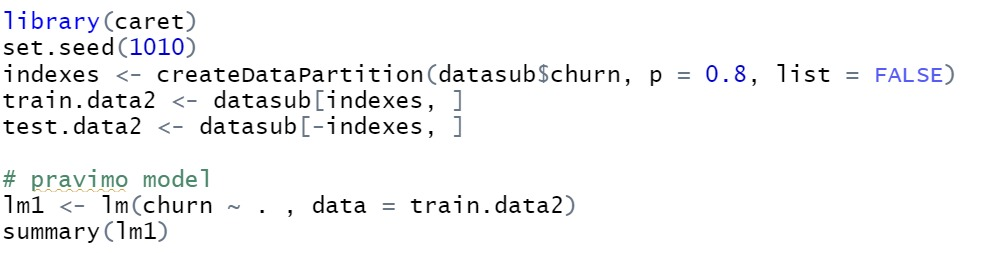
# 4. Linearna regresija

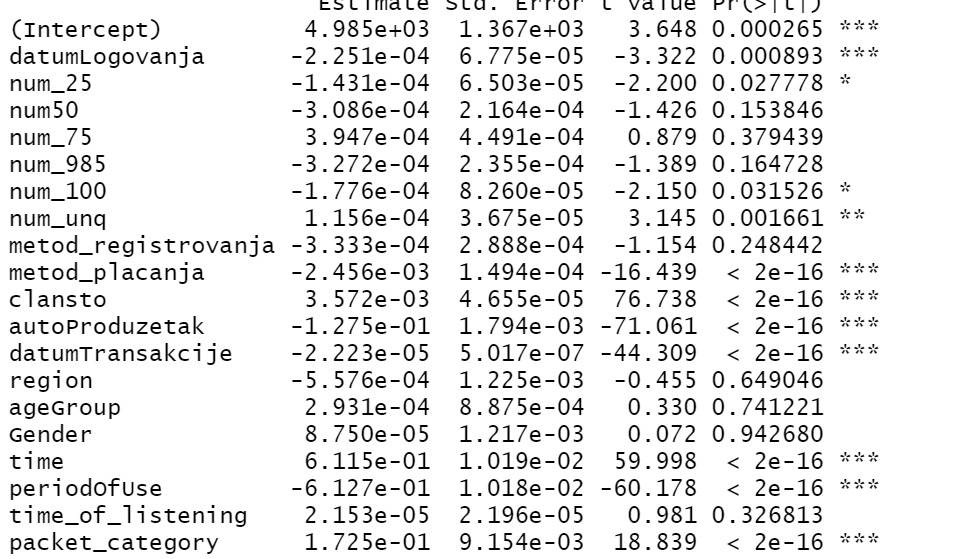
Za linearnu regresiju podaci su sređeni tako da svi, uključujući i izlaznu varijablu budu numeričkog tipa.

Napravljena je korelaciona matrica kako bismo uvideli kolika je korelacija između varijabli.



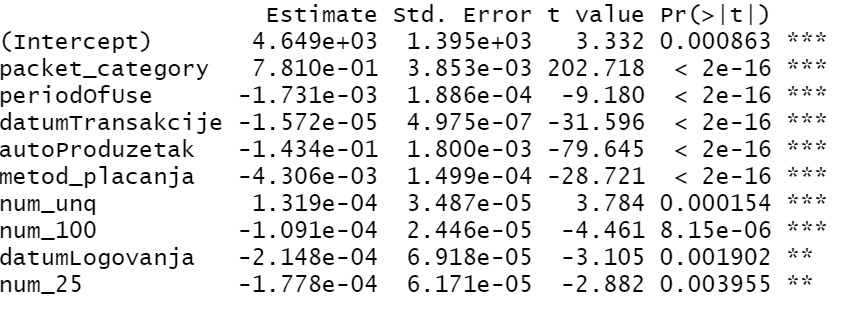
Možemo uočiti da je churn, izlazna varijabla najviše korelisana sa varijablom članstvo ali za pravljanje našeg modela iskoristićemo u početku sve date varijable.

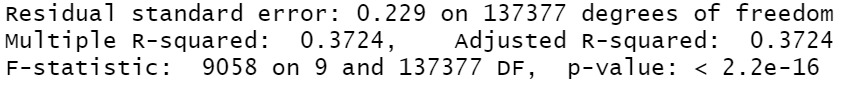




Tri zvezdice pokazuju da su date varijable značajne za predikciju. To govori i t statistika gde ovaj test za nultu hipotezu tvrdi da su varijable nezavisne. Vrednost statistike koja je ispod 0.05 znači da se nulta hipoteza odbacuje i da su varijable zavisne.

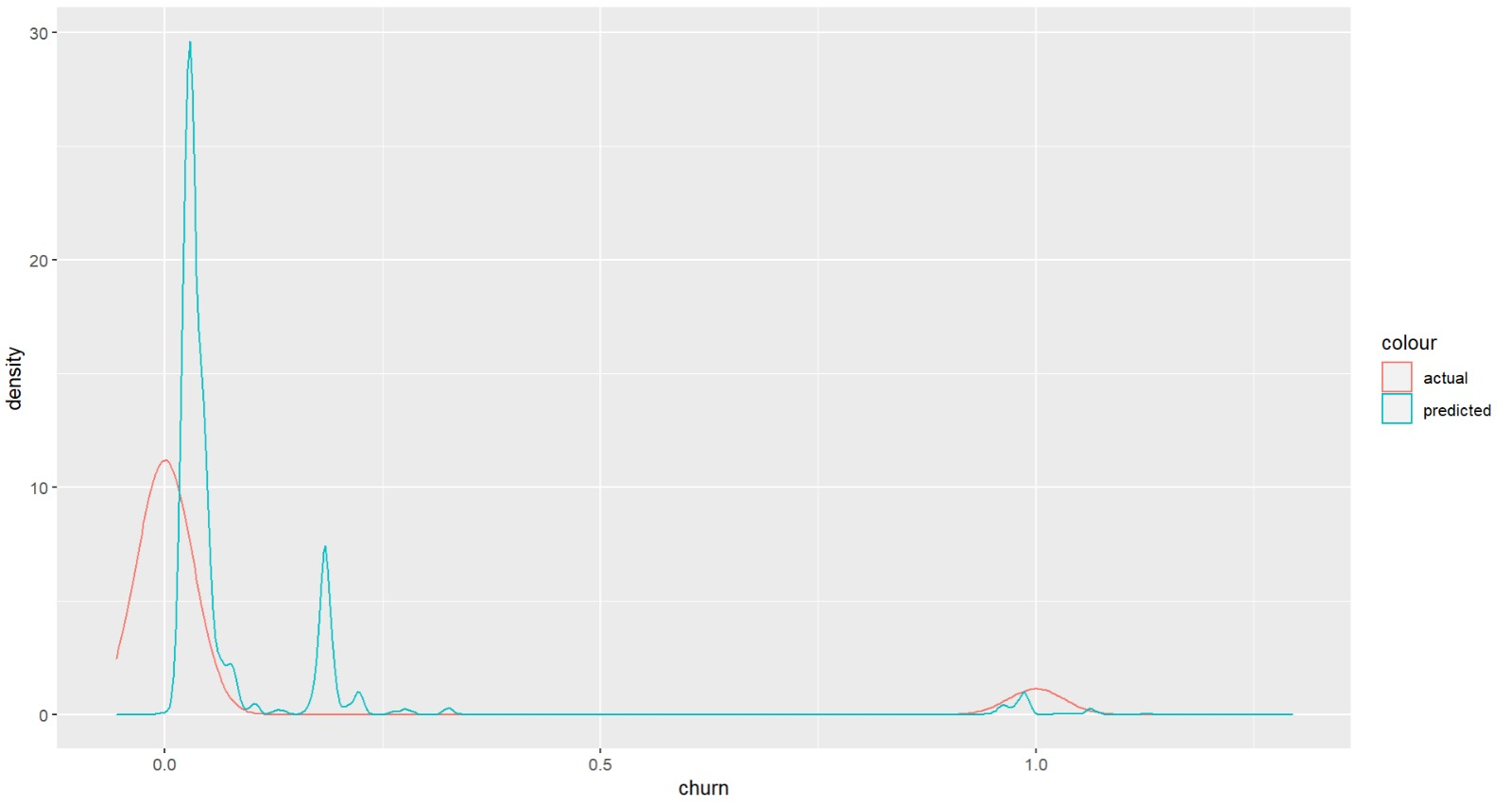
Za sledeći model ćemo uzeti samo zavisne varijable i one koje imaju vif manji od 2, tako da je krajnji model koji de dobija:





Koeficijenti nam govore koliko će se povećati, odnosno smanjiti naša izlazna varijablu u odnosu na jedinično povećanje date varijable.

Rezidual predstavlja razliku između stvarne i predviđene vrednosti. Rsquared govori o tome koliki procenat varijabiliteta ovaj model opisuje. U ovom slučaju to je 37.24 %.

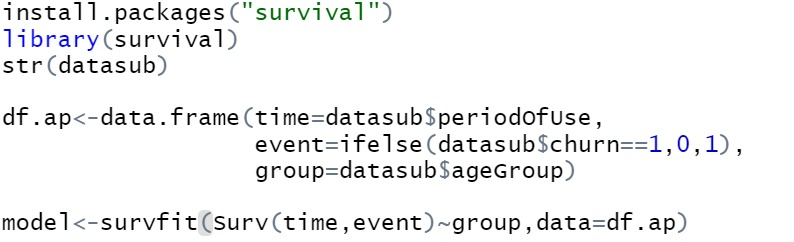


Možemo uočiti da ovaj model nije toliko precizan u predviđanju naše izlazne varijable. Rsquared je 37.2% što je dosta malo I ne možemo uspešno da predvidimo varijabilitete izlazne promenljive. Na slici može da se uoči razlika između stvarne I predviđene vrednosti za izlaznu varijablu churn.

# 5. Analiza preživljavanja

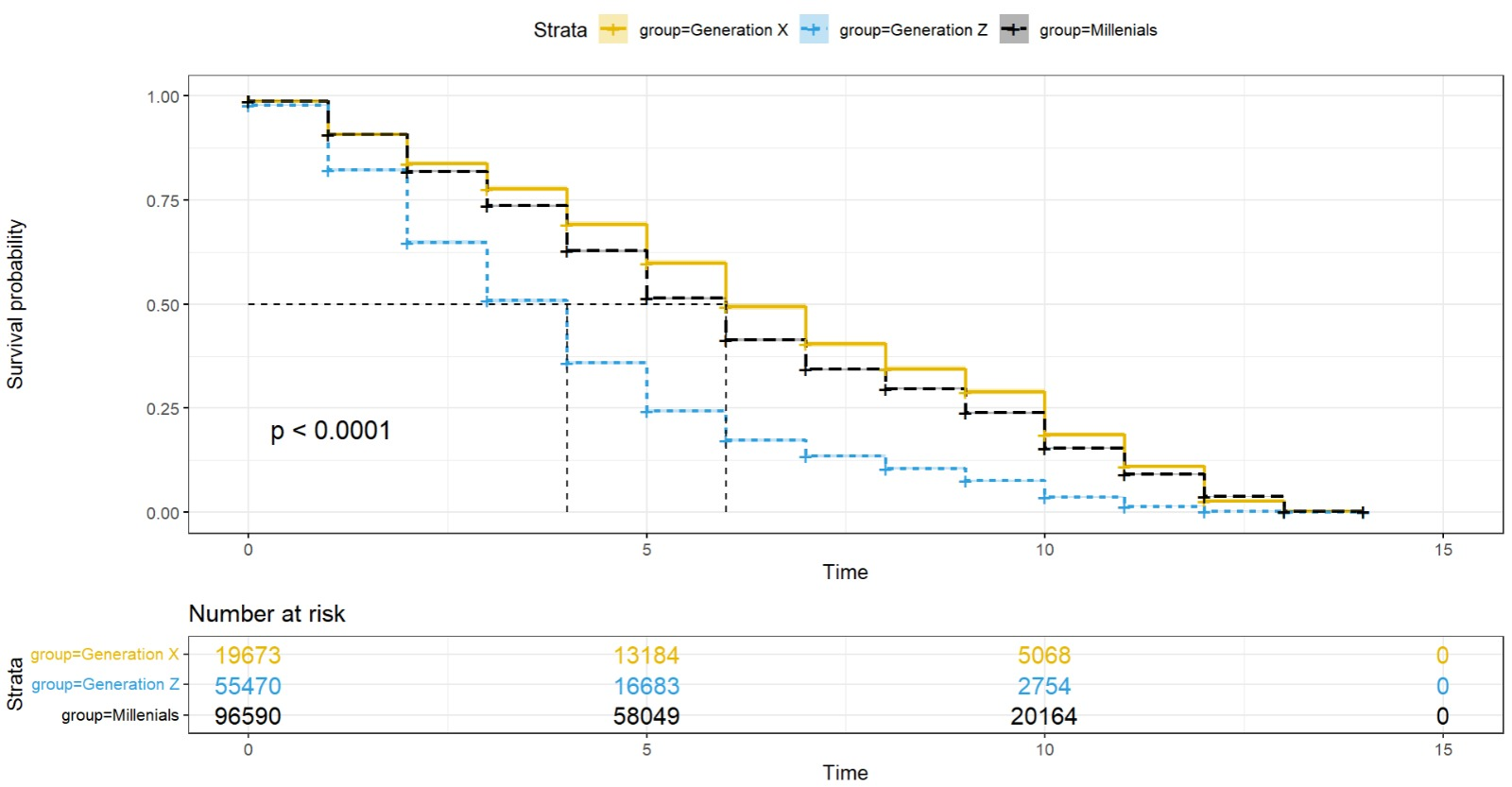
Kroz analizu preživljavanja ispitujemo kolika je privrženost korisnika ovoj platformi. Za promenljivu koja se odnosi na vreme (time) u ovom modelu posmatrali smo varijablu **periodOfUse** koja govori o tome koliko je korisnik proveo na platformi od datuma registracija do datuma kada je prekinuo članstvo. Event predstavlja događaj koji ispitujemo, **churn** u našem slučaju dok ćemo grupisanje izvršiti prema varijabli **AgeGroup** odnosno starosnoj grupi.

Funkcija je:





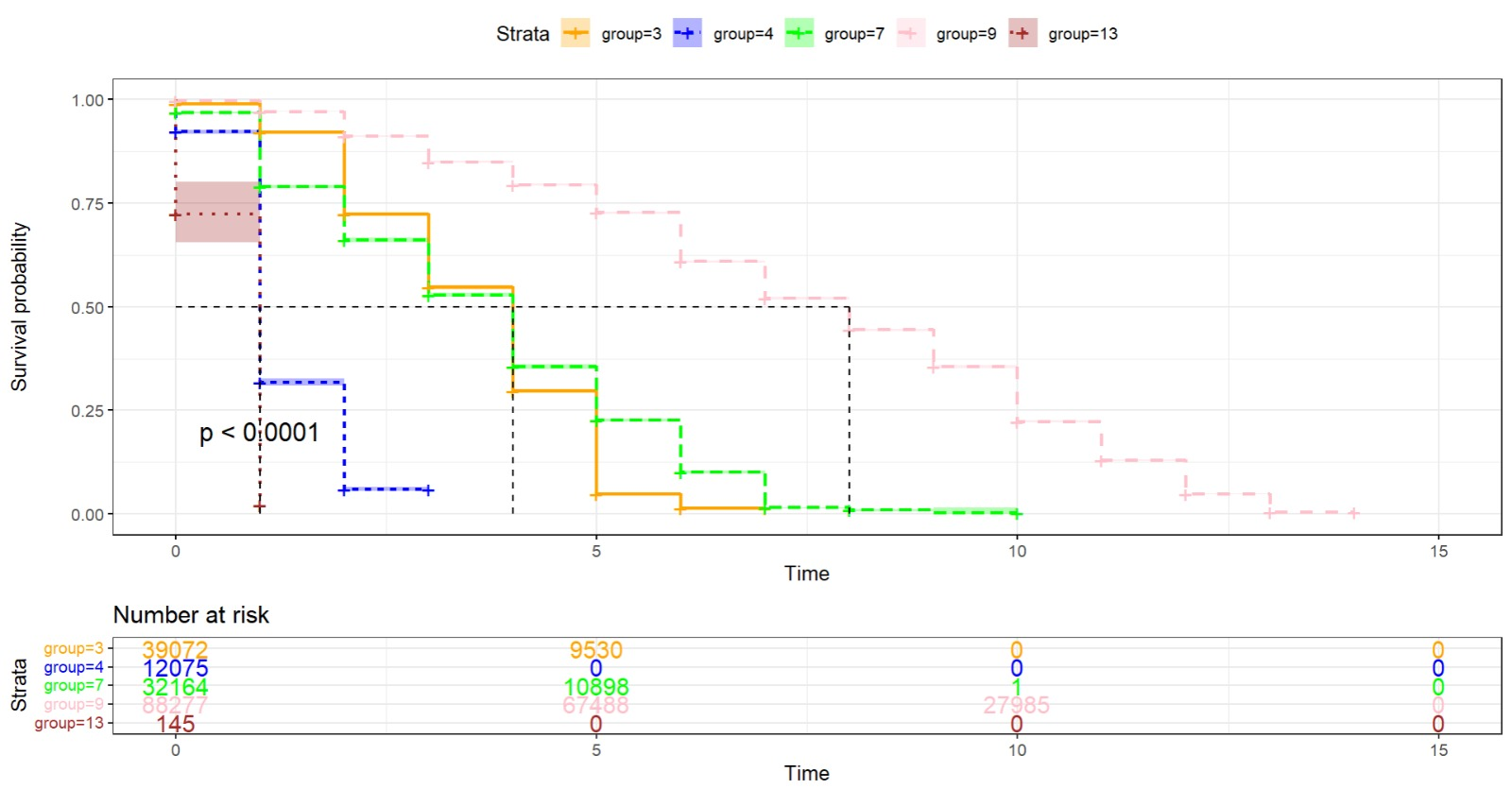
Prikaz analize preživljavanja:



Najveći broj korisnika koji odlazi pripada drugoj grupi, odnosno generaciji Z. Možemo uočiti da 50% ove grupe napušta platformu u prve 4 godine. Dok se sledeće dve grupe značajno poklapaju. Pedeset posto generacije X i Milenijalaca napušta platformu u 6oj godini korišćenja.

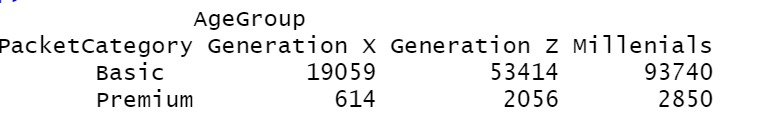
Na osnovu log rank testa vidimo da postoji statistički značajna razlika među ovim grupama jer je p vrednost manja od 0.05.

Sada pravimo model, gde ćemo kategorisati varijable na osnovu metode registrovanja:



Najduže će se zadržati na platformi oni korisnici koji su se registrovali metodom broj 9, dok u najkritičnije ubrajamo grupu 4 i 13.

Na kraju ako poredimo starosne grupe sa paketima koje uzimaju, moženo uočiti da Premium paket uglavnom uzimaju milenijalcia ali i generacija Z ali najveći broj korisnika ukupno pripada grupi milenijalaca.



# 6. Zaključak

Nakon izvršene analize možemo zaključiti da najveći broj korisnika koji odlazi pripada generaciji Z, što ne znači da nisu lojalni brendu jer vidimo da se ovo dešava tek nakon četvrte godine. Razlozi za odlazak ove genracije su mnogobrojni. Zadržavanje mlađe generacije predstavlja veliki izazov za kompaniju zbog većeg broja potreba i pokretljivosti ove generacije.

Najduže na platformi ostaju korisnici koji pripadaju generaciji X, odnosno oni koji imaju preko 40 godina. Nešto manje ostaju i korisnici starosne grupe preko 25 godina (milenijalci). Pored ovoga, posmatrali smo i koji paket pretplate korisnici uzimaju. Paket je podeljen u dve kategorije, premium i basic (manje od 500 n.j.), i zaključili da je najveći broj korisnika koji kupuje premium paketu iz grupe milenijalaca. Time, ako se osvrnemo na pitanje sa početka – Kojoj grupi korisnika kompanija treba da da najviše na značaju, odnosno gde treba da fokusira svoju strategiju zadržavanja kupaca, možemo da zaključimo da bi to bila grupa milenijalaca. U pogledu kriteijuma perioda ostanja na muzičkoj platformi i broja najviše kupljenih premium paketa, milenijalci su najisplativija i najprofitabilnija grupa korisnika za datu muzičku platformu.

Iako je premium paket isplativiji za kompaniju, možemo uočiti da je veći broj korisnika pretplaćen za basic tip paketa. Time opet je preporuka da se kompanija fokusira na milenijalce.

# 7. Reference

[1] Çelik, O., & Osmanoglu, U. O. (2019). Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, *4*(1), 30-38.

[2] Duchemin, R., & Matheus, R. (2021). Forecasting customer churn: Comparing the performance of statistical methods on more than just accuracy. *Journal of Supply Chain Management Science*, *2*(3-4), 115-137.

[3] Barfar, A., Padmanabhan, B., & Hevner, A. (2017). Applying behavioral economics in predictive analytics for B2B churn: Findings from service quality data. Decision Support Systems, 101, 115-127.

[4] Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of marketing research, 43(2), 204-211.