УНИВЕРЗИТЕТ У БЕОГРАДУ
ФАКУЛТЕТ ОРГАНИЗАЦИОНИХ НАУКА

Предмет:
Биостатистика

Пројекат

Тема:
Примена класификационих алгоритама за предвиђање одласка клијената са онлајн стриминг музичке платформе

Студент:

Урош Момчиловић 1030/2015

Милица Јевремовић 257/2015

Београд, јун 2019.

Садржај

1. Идентификација пословног проблема

Проблем постављеног задатка је да прецизно предвидимо да ли ће и који корисници водеће музичке платформе за слушање музике, обновити чланство у наредном месецу или не. Потребно је одредити churn варијабле.

Потребно је креирати предиктивне моделе користећи методе мултиваријационе анализе са циљем креирања најбољег модела за конкретни проблем.

Потребно је средити податке, креирати што већи број модела, уклонити недостајуће вредности и што прецизније решити постављени проблем. При самом коришћењу користимо окружење R studio, који је open-source окружење за развој у R језику. Подаци које је компанија обезбедила налазе се у фајлу: **absChurn.csv.**

Потребно је креирати моделе који ће представљати решења одговарајућих проблема. Креирање предикитивних модела спада у домен предиктивне аналитике. Шта је предиктивна аналитика? Предиктивна аналитика је категорија аналитике података која има за циљ предвиђање будућих исхода на основу историјских података и аналитичких техника као што су статистичко моделирање и машинско обучавање. Наука о предиктивној анализи може генерисати будуће увиде са значајним степеном прецизности. Уз помоћ софистицираних алата и модела за предиктивну аналитику, свака организација може сада да користи претходне и тренутне податке о томе како је прогнозирала трендове и понашања у будућим милисекундама, данима или годинама.

Моћ предиктивне аналитике потиче из широког спектра метода и технологије, укључујући и Биг Дата, копирање података, статистичко моделовање, машинско учење и прикладне математичке процесе. Организације користе предиктивну анализу за праћење тренутних и историјских података како би се на основу добијених параметара открили трендови и прогнозирали догађаји и околности које би требало да се појаве у одређено време.

Са предиктивном аналитиком, организације могу да пронађу и искористе шаблоне који се налазе у подацима како би процениле ризике и могућности. На пример, могу се дизајнирати модели за откривање веза између различитих фактора понашања. Такви модели омогућавале су процену било каквих ризика који су представљали одређену скупну употребу, усмеравајући информацију о доношењу одлуке у категоријама ланца снабдевања и догађаја набавке.

Основна за предвиђање будућих одлука корисника јесу подаци из прошлости. Гледамо податке од купаца који су већ узорковани и њихове карактеристике / понашање (предиктори) пре него што је дошло до напуштања. Прилагођавањем статистичког модела који повезује предикторе са одговором, покушаћемо да предвидимо одговор за постојеће клијенте. Ова метода спада у категорију учења под надзором. У пракси спроводимо следеће кораке како бисмо направили та прецизна предвиђања(Tausend, 2019):

1. Business case: Први корак је заправо разумевање бизниса или употреба случаја са жељеним исходом. Само разумевањем коначног циља можемо изградити модел који је заправо користан. У нашем случају циљ је да се повећа број клијената тако што ће се унапријед идентификовати потенцијални кандидати, и предузети проактивне акције како би их задржали.

2. Прикупљање података и чишћење
   Са разумевањем контекста могуће је идентификовати праве изворе података, чишћење података и припрему за одабир или инжењеринг. Звучи прилично једноставно, али ово је вероватно најтежи део. Модел предвиђања је добар само као извор података. А посебно стартапови или мале компаније често имају проблема да пронађу довољно података да адекватно креирају модел.

3. Избор и инжењерство карактеристика
   Трећим кораком одлучујемо које карактеристике желимо укључити у наш модел и припремити очишћене податке који ће се користити за алгоритам машинског учења за предвиђање одлива корисника.

4. Моделовање. Са припремљеним подацима спремни смо да напунимо наш модел. Али, да бисмо направили добра предвиђања, прво морамо да пронађемо прави модел (селекцију) и друго да проценимо како алгоритам заиста функционише. Иако то обично траје неколико итерација, задржаћемо ово прилично једноставно и зауставити се чим резултати одговарају нашим потребама.

   Последње, али не и најмање важно, морамо проценити и интерпретирати резултате. Шта то значи и које акције можемо извести из резултата? Зато што је предвиђање куповине само пола дела и многи људи заборављају да само предвиђајући могу ићи. У нашем случају ми заправо желимо да их зауставимо.

Churn је присутан у индустрији видео игара. Највећи број новорегистрованих играча напуштају конкретну игрицу само пар дана након регистрације (отварања налога за игру). Фокус је на превенцији одласка. Проблему се приступа из два угла: идентификација раног одласка и превенција истог. За превенцију, прати се понашање самог играча, уочавају се аспекти игре који му се посебно свиђају, и шаљу се обавештења која су прилагођена његовим интересовањима, како би придобили назад датог играча. На овај начин, успешно је редукован churn на 28 %.

Најважнији разлог за улагање у задржавање профитабилних корисника је поређење трошкова привлачења нових корисника и задржавање трошкова за постојеће кориснике. Coopers & Lybrand је показао да је најмање пет пута теже привући нове кориснике од задржавања постојећих (понекад и до 25 пута теже).

Основа за стратегију задржавања клијената треба да буде подстицање дугорочних односа са клијентима кроз поверење, брз одговор на захтеве корисника, висок ниво услуге и поузданост. Кључ за ову стратегију би била способност телекомуникационе компаније да користи поуздане и квалитетне информације о клијентима и да понуди највиши ниво услуге. Такође, стратегија мора бити прилагођена одређеним сегментима.

Фиксне цене, за разлику од реалних које се добијају на основу стварне употребе у неким областима доминирају. Једна од тих области је управо и стриминг музике тако да корисници радије плаћају фиксну цену иако би реална цена била мања. У зависности од величине компаније која пружа ове услуге може се видети какав начин наплаћивања услуга даје најбоље перформансе у односу на профит и одлив корисника. Углавном, мање компаније могу да опстају са фиксним наплаћивањем услуга, док корисници имају већа очекивања од великих компанија тако да је стопа одлива већа.

Миграција корисника је пословни "проблем" којем овакве компаније морају увек да се посвете и усмере на то да буду висококвалитетни и дугорочни учесници на тржишту. Користећи ефикасне CRM технологије, компаније могу на време открити потенцијалне "прекидаче" и предузети превентивне мере како би их задржале као своје клијенте. Помоћу истих метода, најпрофитабилнији сегменти могу се анализирати и посветити посебну пажњу корисницима који, применом одговарајућих програма лојалности за компанију, могу допринијети стварању додатне вриедности и за њих и за компанију.

Када се развије стратегија задржавања корисника, главни циљ је креирање модела који, на основу података из претходних месеци, може предвидети ко ће се пријавити у наредном месецу.

2. Подаци

Бавимо се предвиђањем понашања корисника водеће платформе за онлајн слушање музике. Над датим узорком врши се примена горе алгоритама учења као и анализа постигнутих резултата, како би се на крају, на јасан и концизан начин представила решења описаног проблема.

Сама компанија заинтересована је за разумевање узорка понашања корисника и њихових активности. Подаци о корисницима и њиховом доласку и одласку, о односно престанку коришћења дате услуге су доступни и обрадиви.

Дати скуп података се састоји од информација о томе да ли је корисник престао да се претплаћује на жељени сервис, односно одустао од услуга компаније и ова променљива се назива Churn.

Такође, овај скуп података садржи податке од историји сваког клијента и његовој историји активности на овој платформи. Информације о услузи за коју се клијент пријавио су доступне у скупу података о клијенту.

Поред ових података, могуће је генерисати и географске, као и демографске податке о свим клијентима, старосно доба и пол.

	id	datumLogov	num_25	num50	num_75	num_985	num_100	num_unq	ukupno_vre	grad	godine	pol	metod_regis
1	//0dSjUNUIT	20170327	69	13	6	3	80	117	20851.141	Monako	22	male	9
2	//2cvK2gfq1	20170311	5	4	2	2	12	21	4220.889	Karakas	24	male	9
3	//3f/r1eOH9	20170319	0	0	0	0	4	2	1091.061	Monako	27	male	9
4	//4ANUrAXu	20170316	2	0	0	1	66	66	15616.82	Monako	34	male	9
5	//4hBneqk/	20170306	8	2	3	2	6	18	2851.328	London	19	female	3
6	//5lS6LzJu1s	20170302	5	3	1	0	4	11	1379.249	Male	28	male	9
7	//5vS0wRPz	20170327	0	0	3	0	6	7	2057.053	Nju Delhi	43	female	7
8	//5Ypi+LXhV	20170324	7	0	1	0	73	6	17423.034	Monako	17	female	4
9	//7GpzxjdHC	20170302	0	0	1	1	9	11	2687.467	Monako	31	male	9
10	//8eDSbhxbl	20170312	1	1	0	0	100	96	25947.725	London	18	male	9
11	//aLV1+7Ye6	20170322	3	1	1	0	10	13	2800.359	Oslo	30	male	9
12	//AMgv43wv	20170322	11	11	1	5	4	23	3585.827	Oslo	25	male	9
13	//aQJrbElwe	20170329	60	11	0	8	17	77	7193.799	Male	25	male	9
14	//aWvY7Y+d	20170309	4	3	0	0	8	12	2404.439	Monako	22	female	9
15	//BjtSmLank	20170326	3	2	0	0	225	193	56613.429	Male	18	male	4
16	//cgLNGcURi	20170316	17	4	1	3	57	69	16250.126	Budmipesta	20	male	7
17	//dgmV2cL0(	20170330	0	0	0	0	3	3	706.309	Seul	20	male	7
18	//E3MRiLuu:	20170329	2	1	0	2	22	23	5806.974	Peking	38	female	9
19	//e7ArGzuyp	20170302	1	0	1	0	11	12	2947.809	Oslo	40	female	9
20	//fXDlz34vl4	20170301	28	2	2	0	98	114	25436.88	Monako	20	male	7
21	//HCaBqbZZ	20170316	0	0	0	1	18	15	4409.989	Monako	31	male	9
22	//IgzaMRJo:	20170308	3	1	1	0	5	9	1362.58	Karakas	23	female	9
23	//iqTW5VqN	20170306	7	3	0	1	8	19	2390.573	Seul	39	male	9
24	//jEIR1zQa7	20170330	5	0	1	2	23	22	6570.656	Oslo	24	female	9
25	//JrJATHAxT:	20170313	4	4	1	0	41	32	9403.579	Nju Delhi	25	male	3
26	//LE/G/D+EV	20170320	2	0	0	0	5	4	1167.74	Karakas	25	male	3
27	//LSJMhihqC	20170321	13	2	3	2	20	33	6395.359	Monako	25	male	7
28	//lTUsQvE+U	20170302	3	3	0	0	30	24	7695.021	Sangaj	33	male	9
29	//lyN+S5Ruj	20170331	1	0	1	0	26	20	6240.691	Karakas	59	male	9

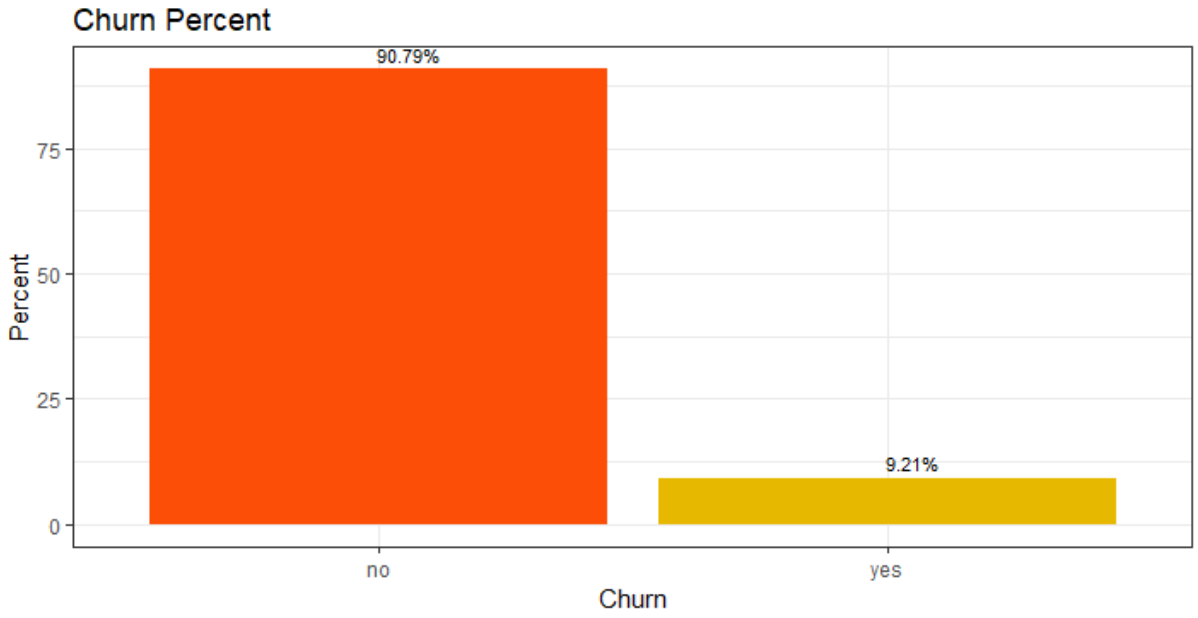
Слика 1. Преглед података

На самом почетку рада, скуп података обухватао је 171733 опсервација (клијената) описаних помоћу 23. атрибута (варијабли) приказаних на слици 1. Када је реч о типовима података, промењиве пол и град су текстуалне, променљива која означава укупно време слушања песама изражено у секундама је нумеричког типа, док су све остале променњиве целобројног типа, што се јасно може видети из прегледа података.

2.1 Експлораторна анализа података

1.1.1. Визуелизација података

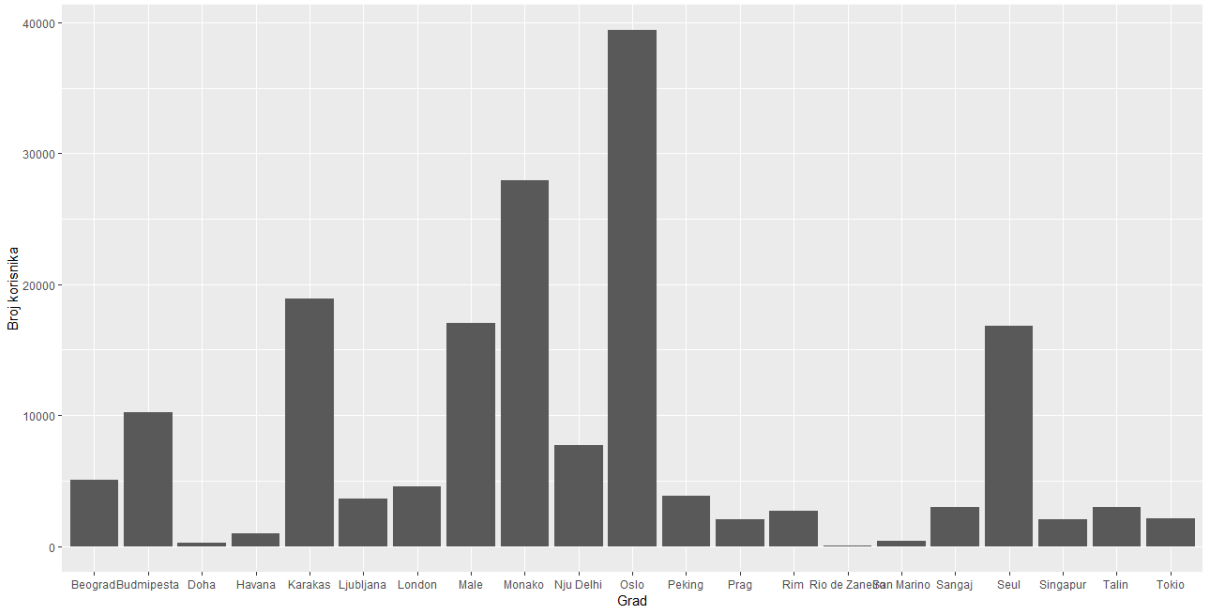
У датим подацима имамо податке и о churn променљивој, односно о укупном броју одлазака клијената и тај однос је приказан на следећем графикону.



Графикон 1. Процентуални однос клијената у односу на варијаблу churn

На почетку визуелизације података, обратићемо пажњу на churn у односу на променљиву ауто продужетак.

Уочава се да је број клијената који су се одлучили за опцију ауто продужетка услуге и који су одлучили да раскину уговор са музичком платформом доста мањи од броја клијената који су раскинули уговор са напоменутом платформом, а нису се иницијално одлучили за опцију ауто продужетка. Једно од тумачења овако логичног исхода јесте да људи који су изабрали опцију ауто продужетка вероватно имају протходно позитивно искуство са датом платформом и немају намеру да у скоријем временском периоду прекину коришћење дате услуге. Са друге стране, клијенти који се нису одлучили за ауто продужетак, вероватно желе на кратак временски период да опробају услуге платформе, што доводи то тога да су неки клијенти задовољни пруженом услугом а неки не.



Графикон 2. Број корисника услуге у односу на град у коме живе

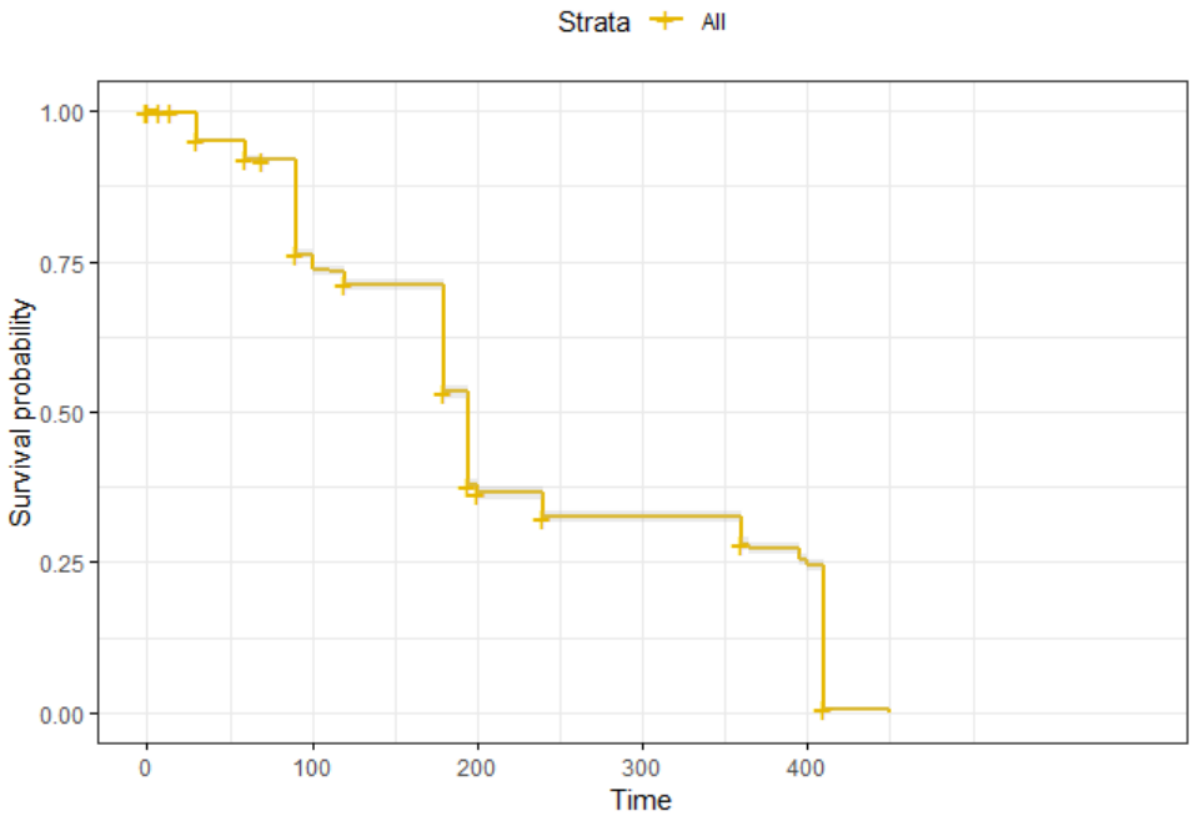
Са графикона 2. уочавамо да већина корисника услуге дате музичке платформе живи у развијеним градовима и окружењима. Тумачење оваквог резултата може бити да напоменута платформа као циљно тржиште првенствено погађа економски платежно становништво.

3. Анализа преживљавања

Проблем постављеног задатка је да прецизно предвидимо да ли ће и који корисници водеће музичке платформе за слушање музике, обновити чланство у наредном месецу или не. Потребно је одредити churn варијабле и вероватноћу преживљавања. У модел убацујемо clansto и churn варијабле из датасета.

sfit<-(survfit(Surv(clansto, churn)~ 1, data = datas))

summary(sfit)$table



Графикон 5: Вероватноћа наредне претплате у односу на дужину коришћења платформе

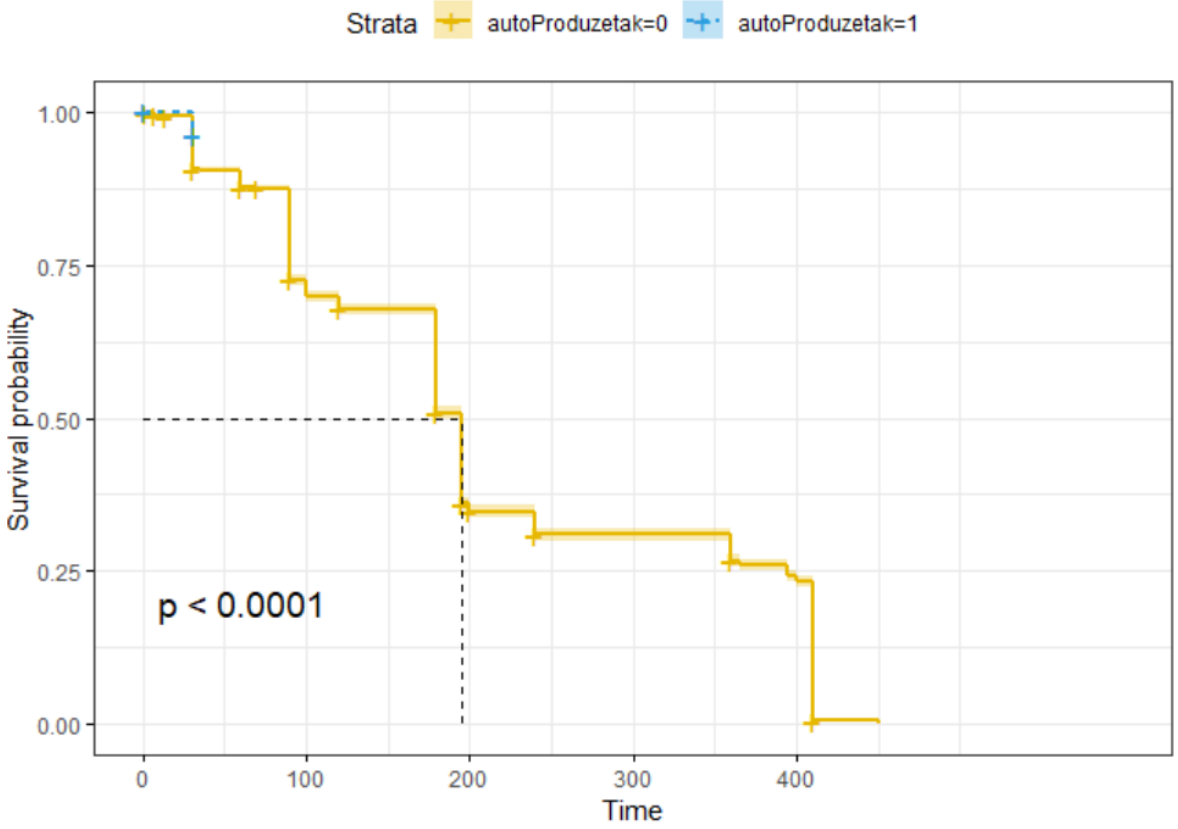
Позивом summary(sfit)$table, добијамо увид у медиану времена „преживљавања“ корисника платформе у данима ка churn варијабли и она износи 195 дана. То се може видети на слици испод.

Каплан-Мајер график на слици испод може се тумачити на следећи начин: Хоризонтална оса (x-оса) представља чланство у данима на платформи, а вертикална оса (y-оса) показује вероватноћу преживљавања и претплате у наредном времену. Линије представљају криве преживљавања групе. Вертикални пад у кривинама означава догађај.

1. У времену нула, вероватноћа преживљавања је 1.0 (100% корисника на платформи).

2. У времену 190, вероватноћа преживљавања је око 0,50 (или 50%) за аутоПродужетак.

3. Медијана преживљавања је 195 коју можемо видети позивом функције *summary(fit) $table* која је разјашњена у скрипти.



Графикон 6: Каплан Мајер крива варијабле (0-1) аутопродуживање

Након извршавања *survdiff, surv_summary()* функција, како би имали увид у резултате примењујемо и Cox PH регресију:

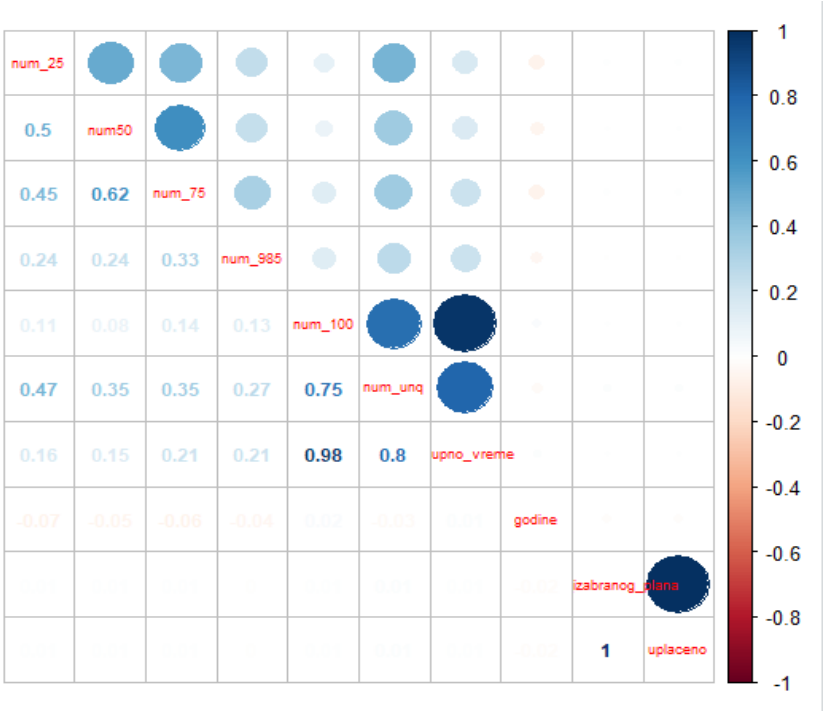
fit <- coxph(Surv(clansto, churn) ~ autoProduzetak, data = datas)

Након позива функције резултати су приказани у скрипти. Овде ћемо протумачити део резултата. Exp(coef) представља однос ризика - мултипликативни ефекат те варијабле на стопу опасности за сваку јединицу се повећава променљива. Колона exp (coef) је однос ризика, тј. можемо тврдити да за варијаблу продужетка (0-1) око 60% можемо бити сигурни за смањени ризик. Дакле, закључак је да ће продужити претплату за следећи месец 0.387% корисника. За варијаблу као што је „претплата“, иде од не-претплате 0 (основна) до резултата који резултирају да је за претплату приближно 60% смањења “штета преживљавања“. Исто тако, можемо ставити знак на coef и узети exp (0.94693), што можемо тумачити као не-претплату, 0, што је резултирало 0.02 пута повећање ризика, или да се избор не-претплате дешава отприлике за стопу од 0.02 по јединици времена.

4. Приступ и методологија

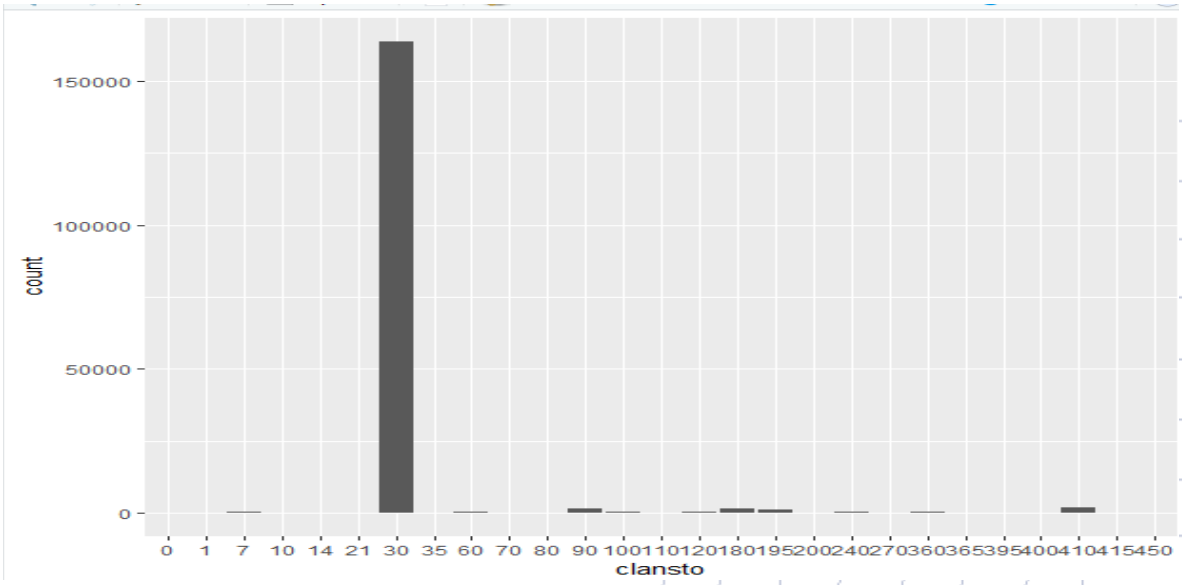
Почели смо са дефинисањем које све варијабле имају информативну вредност. У првом кораку смо избацили све идентификаторе.

Након тока смо податке раздвојили на нумеричке и категоријске. За нумеричке варијабле смо проверили коефицјенте корелације. Варијабле које су међусобно високо корелисане избацили смо из даље анализе.



Графикон 3

За категоријске варијабле проверили смо рапсоделу вредности по класама. Оне код којих постоји дизбаланс у расподели избацили смо из даље анализе.



Графикон 4

Проверили смо недостајуће вредности којих није било.

Овако припремљене податке поделили смо на тренинг и тест скуп у односу 70 према 30.

5. Литература

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*

Andy Field, Jeremy Miles(2016). *Discovering Statistics Using R*

Martins, H. (2017). *Predicting user churn on streaming services using recurrent neural networks.* Stockholm: Royal institute of technology, School of computer science and communication.

S. Christian Albright, Wayne L. Winston(2010). *Business Analytics: Data Analysis & Decision Making*

https://towardsdatascience.com/churn-prediction-770d6cb582a5