УНИВЕРЗИТЕТ У БЕОГРАДУ

ФАКУЛТЕТ ОРГАНИЗАЦИОНИХ НАУКА

ЗАВРШНИ РАД

Тема: Примена метода рачунарске интелигенције за предвиђање лојалности клијената

Ментор:

др Ивана Драговић

Студент

Алекса Стефановић

160/15

Београд, 2019. године

Примена метода рачунарске интелигенције за предвиђање лојалности клијената

Апстракт

Предвиђање понашања корисника представља значајан задатак већине компанија. Међу свим индустријама које се баве овим проблемом, област телекомуникација се налази на врху листе са приближном годишњом стопом раста од око 30%. Постоје различити приступи приликом решавања проблема одлазака корисника кроз развијање различитих предиктивних модела. Поред дискриминационе анализе, како линеарне тако и квадратне, овај рад обухвата још неке алгоритме класификације као што су: логистичка регресија, метода потпорних вектора(енгл. *Support Vector Machine*) и К најближих суседа(енгл. *K-Nearest Neighbour*). Такође, крајњи циљ подразумева дефинисање модела за идентификацију корисника за које постоји велика вероватноћа да ће отказати претплату на телефонску услугу у наредних 30 дана. За вредновање резултата алгоритама коришћене су евалуационе метрике које описују њихову прецизност, тачност, одзив и Ф-1 скор. На крају рада, дат је табеларни приказ самих резултата који су добијени на основу појединачних алгоритама као и закључак у коме су дати даљи правци истраживања.

Кључне речи: *churn*, машинско учење, метода потпорних вектора, К најближих суседа, дискриминациона анализа, логистичка регресија.

Садржај:

Списак слика и графикона

Списак табела

1. Увод

Привлачење нових и задржавање старих корисника један је од најзначајнијих задатака компанија. Док се нове компаније концентришу на стицање нових корисника, зреле се фокусирају на задржавање постојећих како би себи пружиле могућност унакрсне продаје (уз један производ/услугу купци узимају још један сродан производ/услугу).

Предвиђање корисника или предвиђање понашања корисника представља проблем идентификације корисника који ће вероватно престати да користе одређен производ или услугу. Компаније за пружање телефонских услуга, интернет провајдери, компаније које се баве пословима осигурања и мониторинга, често користе ову анализу као једну од својих главних пословних метрика јер су трошкови задржавања постојећих корисника далеко мањи од стицања нових.

Конкретно, разматра се питање идентификације корисника који могу напустити компанију (енгл. *customer churn*) и самим тим престати користити њен производ односно услугу. Управо овакве информације могу бити од велике вредности за компаније у напорима да задрже постојеће кориснике. Телеком и банкарски сектор су, од самог почетка, највише заинтересовани управо за овакве информације, а у последње време те информације добијају на значају када се у разматрање узме све интензивнија конкуренција која отвара нова тржишта. Са друге стране, у ери онлајн и дигиталних платформи, покушаји предвиђања одлазака корисника из великих скупова података постали су релевантни за широк спектар услуга; *streaming* дигиталних медија као што су музика и видео садржај, видео игре, итд.

Суочене са овом претњом, компаније би требало да буду опремљене најефикаснијим и најефективнијим методама за испитивање понашања својих корисника, предвиђајући њихов будући корак.

1.1. Дефиниција проблема

Углавном, већина дате литературе овакав проблем, проблем предвиђања понашања корисника, формулише као бинарни класификациони проблем(односно излазна променљива има логичке вредности 0 или 1), што је и најчешћи случај у пракси, па у складу са тим је и дата дефиниција самог проблема:

Потребно је идентификовати кориснике који тренутно плаћају услугу најмање 30 дана и, уз помоћ скорашњих историјских података, предвидети могућност да ће корисници одлучити да прекину коришћење услуге у року од 30 дана. Поред тога, користећи ове информације, идентификовати скуп корисника на које компанија треба више да обрати пажњу и самим тим искористити такве предности у маркетиншкој кампањи.

Претпоставка је да је 30 дана најмање потребно време да се прикупи довољна количина података, али и да се осигура да нови корисници могу бити део скупа података. Историјски подаци јесу друго ограничење датог проблема са којима се компаније суочавају у смислу рачунских ресурса за прикупљање података и агрегацију истих.

1.2. Циљ рада

Циљ овог рада је да се спроведу и касније упореде различити алгоритми из области машинског учења да би се пронашао онај који даје најбоље резултате при идентификацији корисника који ће вероватно активно отказати плаћену претплату.

Начином идентификовања корисника који су спремни да напусте плаћени пакет, компанија може покренути циљане маркетиншке кампање и можда убедити део њих да остану нудећи им бољу понуду или друге опције јер је свакако задржавање старијих корисника знатно јефтиније од стицања нових. Да ли би такви покушаји били успешни или не, то је друго питање, али без довољно ефикасног модела предвиђања одласка, покушај задржавања корисника био би тешко изводљив и значајно би повећао трошкове.

Са методолошког аспекта, рад подразумева примену и каснију упоредну анализу резултата различитих алгоритама за класификацију као што су дискриминациона анализа(линеарна и квадратна), логистичка регресија, метода потпорних вектора(енгл.*Support Vector Machine*) и К најближих суседа(енгл.*K-Nearest Neighbour*).

1.3. Структура рада

Рад започиње дефинисањем истраживачког проблема и пружа увид у његову важност. Други део рада обухвата теоријске основе методологија односно алгоритама који су коришћени. У следећем поглављу описан је скуп података који је коришћен и детаљно су описани одабрани атрибути који имају највећу предикторску моћ. У наставку рада извршено је вредновање и анализа добијених резултата док крај рада чини закључак који се може донети на основу добијених резултата. Такође, у прилогу је дат код, који је имплементиран у развојном окружењу *Rstudio* програмског језика *R*.

2. Теоријске основе

2.1. Машинско учење

Машинско учење је грана вештачке интелигенције заснована на претпоставци да систем може да учи из података односно да идентификује “шаблоне” и законитости у подацима, и да на основу њих доноси одлуке, уз минималну људску интервенцију(Inc., Sas Institute, 2018). Машинско учење настало је из идеје да рачунари могу да уче на основу датих великих скупова података, без претходног експлицитног програмирања, односно дефинисања програмске логике. Итеративни аспект машинског учења допринео је реализацији ове идеје(Inc., Sas Institute, 2018). Основна идеја машинског учења је да креирани алгоритам прима улазне податке и на основу њих, уз помоћ математичких и статистичких анализа и прорачуна, предвиђа излазе. Сваки пут када се унесу нови улазни подаци, алгоритам ажурира предвиђање излаза(Rouse, 2018).

Захваљујући константној изложености великој количини података из разних извора и све повољнијем складиштењу велике количине података, организације свих врста увиделе су потенцијал таквих података. Доступност података, с једне стране, и технолошки напредак

рачунара који отвара могућност да се све веће количине података обрађују брже и лакше, с друге стране, отворили су могућност за широку примену машинског учење у свим областима савременог живота(Inc., Sas Institute, 2018). Само неке од области у којима је машинско учење пронашло своју примену су: банкарство и финансије, маркетинг и продаја, саобраћај, људски ресурси и још много других.

Неки од најпопуларнијих типова машинског учења су надгледано учење(енгл. *Supervised learning*), ненадгледано учење(енгл. *Unsupervised learning*), полунадгледано учење(енгл. *Semisupervised learning*) и учење са подстицајем(енгл. *Reinforcement learning*)(Inc., Sas Institute, 2018).

Овај рад се бави искључиво надгледаним учењем које подразумева да алгоритми учења за сваку опсервацију имају познату излазну вредност и скуп атрибута који описују карактеристике опсервације.

2.2. Класификација

У машинском учењу као и у статистици, класификација представља проблем идентификације категорије(суб-популације) којој нова опсервација припада, на основу тренинг скуп података за чије опсервације је већ познато којим категоријама припадају (James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013).

У раду су коришћена четири модела класификације: логистичка регресија(енгл. *Logistic Regression*), дискриминациона анализа(квадратна и линеарна), К најближих суседа(енгл. *K-Nearest Neighbour*) и метода потпорних вектора(енгл. *Support Vector Machine*).

2.2.1. Логистичка регресија

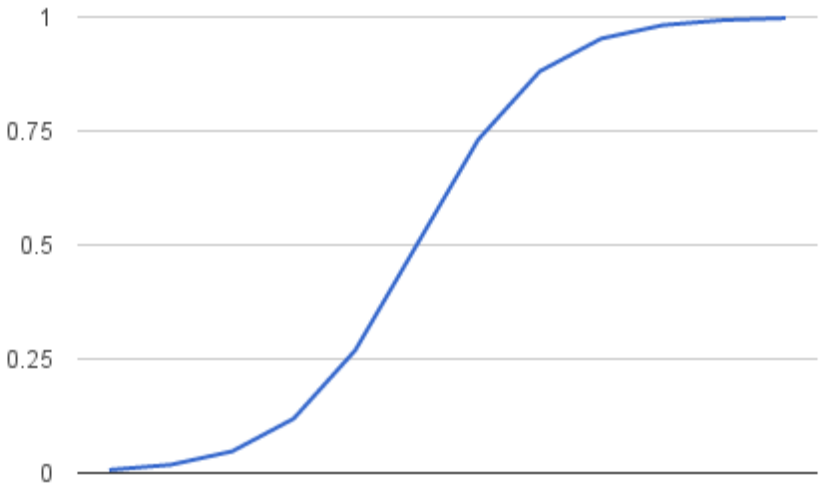
Логистичка регресија је модел класификације који се може брзо обучити и изабран је као основа за поређење са напреднијим моделима.

То је статистичка метода која моделује(израчунава) условну вероватноћу да дата опсервација припадне позитивној класи(*Y = 1*) при датим улазним променљивим, **x**, дефинисана једначином 2.1. - обично се назива логистичка функција, где су **w** параметри за тренирање модела(James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013).

$$P(Y = 1|\boldsymbol{x}) = \boldsymbol{\sigma}(\boldsymbol{w^T x}) = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e^{-w^t x}}} \qquad 2.1.$$

Како би се добили оптимизовани параметри, користи се максимална вероватноћа. Једначина 2.2. приказује функцију губитка која је дефинисана узимањем негативног логаритма вероватноћа где је *N* укупан број предвиђања, $y \in \{0, 1\}$, а $p_n = \boldsymbol{\sigma}(\boldsymbol{w^T x})$ је предвиђање.

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N}\{y_n \ln p_n + (1 - y_n)\ln(1 - p_n)\} \qquad 2.2.$$



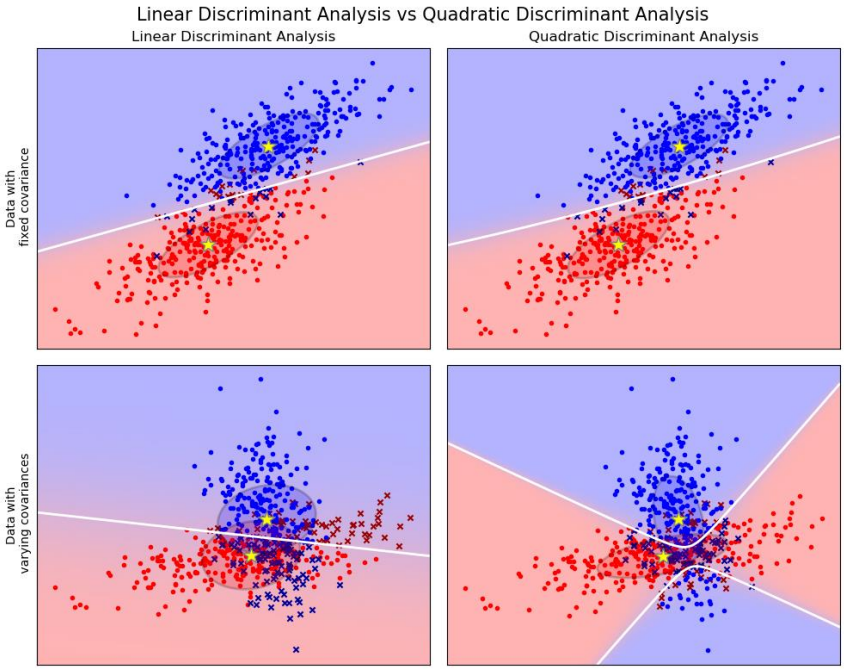
***Слика 1**: Функција логистичке регресије*[1]

Са аспекта предвиђања, потребно је знати у којој од две могуће групе спада свака опсервација. Преко логистичке регресије добијају се решења код којих ће зависна променљива имати вредност негде између 0 и 1. Предвиђена вредност је вероватноћа да ће јединица посматрања припасти једној или другој групи, што је управо оно што је потребно приликом решавања овог проблема(Walker SH, Duncan DB, 1967).

2.2.2. Дискриминациона анализа

Линеарна дискриминациона анализа представља методу која се користи у статистици, препознавању узорака и машинском учењу ради проналажења линеарне комбинације карактеристика која карактерише или раздваја две или више класа предмета или догађаја (Gareth M. James, Trevor J. Hastie, 2002). Добијена комбинација може се користити као линеарни класификатор или, чешће, за смањење димензија пре саме класификације.

[1] https://machinelearningmastery.com/logistic-regression-for-machine-learning/



***Слика 2**: Линеарна дискриминациона анализа*[2]

Циљ линеарне дискриминационе анализе је пројектовање карактеристика у простору више димензије на простор нижих димензија.

То се може постићи у три корака. Први корак је израчунавање одвојивости између различитих класа(тј. удаљеност између средње вредности различитих класа) које се такође назива и варијанса између класа:

$$S_b = \sum_{i=1}^{g} N_i(\bar{x}_\iota - \bar{x}) - (\bar{x}_\iota - \bar{x})^T \qquad 2.3.$$

Други корак је израчунавање удаљености између средње вредности и узорка сваке класе, што представља варијансу унутар класе:

$$S_w = \sum_{i=1}^{g}(N_i - 1)S_i = \sum_{i=1}^{g}\sum_{j=1}^{N_i}\left(x_{i,j} - \bar{x}_\iota\right)\left(x_{i,j} - \bar{x}_\iota\right)^T \qquad 2.4.$$

Трећи корак је изградња простора нижих димензија који максимизира варијансу између класе и минимизира варијансу унутар класе. П је пројекција простора у нижој димензији, која се назива Фишеров критеријум(Gareth M. James, Trevor J. Hastie, 2002).

[2] https://scikit-learn.org/stable/modules/lda_qda.html

$$P_{lda} = arg_p \max \frac{P^T S_b P}{P^T S_w P} \qquad 2.5.$$

Квадратна дискриминациона анализа подразумева да свака класа користи сопствену процену варијансе(или коваријансе када постоји више улазних променљивих)(Gareth M. James, Trevor J. Hastie, 2002).

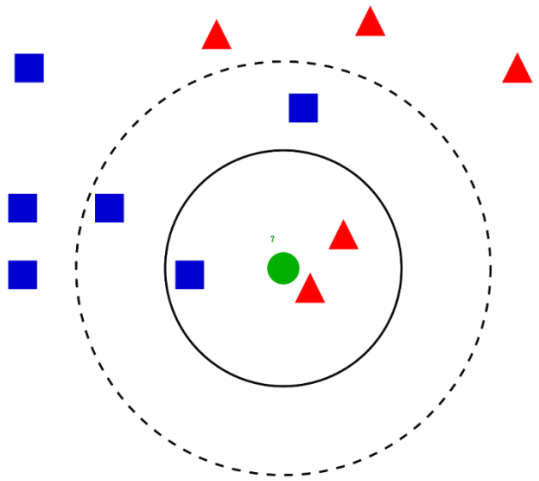
2.2.3. К најближих суседа(K-Nearest Neighbour)

К најближих суседа је непараметарски алгоритам за класификацију познат и као "*lazy learning*" алгоритам. Непараметарски значи да не даје никакве претпоставке о основној дистрибуцији података. Стога КНН треба да буде један од првих избора када је у питању класификациони проблем односно када се мало зна о дистрибуцији самих података. "*Лењ*" алгоритам значи да не користи податке о тренингу односно не постоји експлицитна фаза тренинга. У КНН-у је дата тачка података класификована на основу класе најближих К комшија. К је обично непаран број у случају бинарне класификације. У случају проналажења најближих комшија најчешће се користи еуклидско растојање(Anton, Howard, 1994):

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \qquad 2.6.$$

Кораци су следећи:

1. Израчунање удаљености између сваке инстанце,
2. Проналажење најближих комшија,
3. Бирање оне класе где је већина суседа.



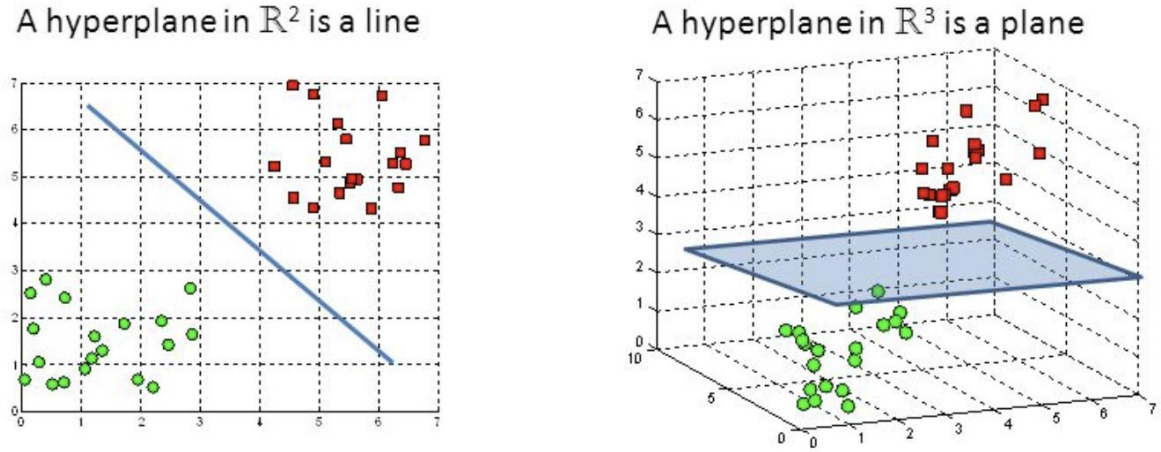
***Слика 3**: Начин функционисања knn алгоритма*[3]

[3] https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

KHH се најчешће користити за класификацију: излаз је припадност класи(предвиђа класу - дискретну вредност). Објекат је класификован већином гласова својих суседа, при чему је објекат додељен класи која је највише заступљена међу својим најближим суседима. Може се користити и за регресију: излаз је вредност објекта(предвиђа нумеричке вредности). Ова вредност је просек(или средња вредност) вредности њених најближих суседа(Onel Harrison, 2018).

2.2.4. Метода потпорних вектора(Support Vector Machine)

Овај алгоритам, слично као и претходни, се може користити за решавање проблема како класификационих тако и регресионих. Циљ методе потпорних(носећих) вектора јесте да се пронађе хиперраван у *n*-димензионалном простору, при чему *n* представља број атрибута, који јасно разврставају тачке појединих елеманата или инстанце. За одвајање две класе инстанци, може се изабрати велики број различитих равни. Циљ је пронаћи раван која има максималну маргину, тј. максималну удаљеност између инстанци обе класе. Максимизирање маргине омогућава да се будуће инстанце могу класификовати са већом поузданошћу(James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013).



***Слика 4**: Приказ хиперравни које раздвајају податке различитих класа*[4]

У случају да постоји линеарна хиперраван која може да подели инстанце на два дела користи се линеарни тип овог алгоритма:

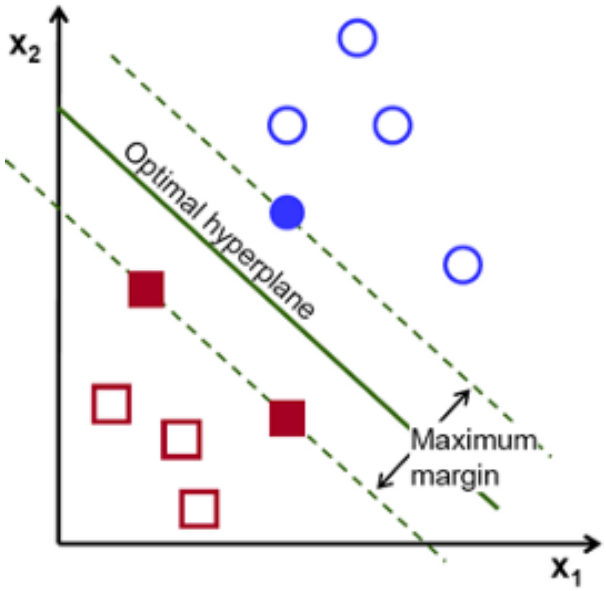
$$\vec{w} * \vec{x} - b = 0 \qquad 2.7.$$

где $\vec{w}$ представља вектор који је нормалан на раван(Cortes Corinna, Vapnik, Vladimir N., 1995).

Уколико је тренинг скуп података линеарно дељив могуће је дефинисати две паралелне равни које раздвајају две класа података, тако да је растојање између њих максимално.

[4] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Део који је ограничен овим равнима назива се маргина, а раван која максимизира маргину јесте раван која која се налази у средини паралелних равни(Hsu, Chih-Wei, Chang, Chih-Chung & Lin, Chih-Jen, 2003).



***Слика 5**: Хиперраван која максимизира маргину*[5]

Такве равни дефинисане су изразима:

$$\vec{w} * \vec{x} - b = 1 \qquad 2.8.$$

све што је на или изнад ове равни припада првој класи,

$$\vec{w} * \vec{x} - b = -1 \qquad 2.9.$$

све што је на или испод ове равни припада другој класи.

Предности ове методе су:

- Ефикасност у вишедимензионалним просторима,
- Користи различите функције за различите проблеме одлучивања,
- Могућност креирања комбинација различитих функција.

Недостаци:

- Слабе перформансе када је број атрибута већи од самог узорка,
- Не обезбеђује процене вероватноће.

Међутим, данас су све чешћи проблеми код којих скупови података нису линеарно дељиви тако да се у том случају дефинишу различите друге маргине као што је "*soft-margin*". Такви

[5] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

проблеми се решавају коришћењем и вишедимензионалних простора. Како ово није тема самог рада, овај алгоритам неће бити детаљније описан.

3. Подаци

У конкретном раду, испитиваће се предвиђање понашања корисника телекомуникационе компаније, над подацима које је изложила заједница *IBM Watson Analytics* компаније у оквиру експертског блога(Stacker, 2015). Узорак који је објављен може се наћи у прилогу овог рада, а у наставку је дат преглед скупа података(енгл. *data set*), над којим ће се вршити примена горе поменутих алгоритама машинског учења као и анализа постигнутих резултата, како би се на крају, на јасан и концизан начин донео закључак о могућностима решавања овог проблема. Приликом разматрања овог проблема укључени су само они атрибути који имају највећу предикторску моћ.

3.1. Опис података

Телекомуникациона компанија изложила је податке о својим корисницима и њиховом престанку коришћења дате услуге. Заинтересована је за разумевање узрока понашања корисника и њихових активности како би се идентификовали корисници на које је потребно обратити пажњу.

Скуп података се састоји од следећих атрибута:

- ***customerID*** – идентификациони број корисника
- ***gender*** – пол корисника(мушки, женски)
- ***SeniorCitizen*** – да ли је корисник старији или не при чему постоји граница која је дефинисана унутар саме компаније(1, 0)
- ***Partner*** –да ли корисник има партнера или не (Да, Не)
- ***Dependents –*** да ли имају децу или не(Да, Не)
- ***tenure –*** број месеци које је корисник провео у компанији
- ***PhoneService*** – да ли поседује телефон или не (Да, Не)
- ***MultipleLines*** – да ли поседује више телефонских линија (Да, Не, Нема телефонску линију)
- ***InternetService*** – интернет сервис провајдер(*DSL*,*Fiber optic*, Не)
- ***OnlineSecurity*** – да ли поседује онлајн безбедност(Да, Не, Нема интернет)
- ***OnlineBackup*** – да ли има могућност повратка података(Да, Не, Нема интернет)
- ***DeviceProtection*** – да ли постоји могућност заштите уређаја(Да, Не, Нема интернет)
- ***TechSupport*** – да ли има техничку подршку(Да, Не, Нема интернет)
- ***StreamingTV*** – да ли постоји могућност ТВ стриминга(Да, Не, Нема интернет)
- ***StreamingMovies*** - да ли постоји могућност стримовања филмова(Да, Не, Нема интернет)
- ***Contract*** – тип уговора(Месечно, Годину дана, Две године)
- ***PaperlessBilling*** – наплата без папира(Да,Не)

- ***PaymentMethod*** – начин плаћања(Електронски чекови, Пошта, Банкарски трансфери(аутоматски), Кредитне картице(аутоматски))
- ***MonthlyCharges*** – месечна плаћања корисника
- ***TotalCharges*** – укупна плаћања појединачног корисника
- ***Churn*** – да ли је корисник престао да користи услугу или не(Да,Не)

gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
Female	No	Yes	No	1	No	No phone service	DSL	No
Male	No	No	No	34	Yes	No	DSL	Yes
Male	No	No	No	2	Yes	No	DSL	Yes
Male	No	No	No	45	No	No phone service	DSL	Yes
Female	No	No	No	2	Yes	No	Fiber optic	No
Female	No	No	No	8	Yes	Yes	Fiber optic	No
Male	No	No	Yes	22	Yes	Yes	Fiber optic	No
Female	No	No	No	10	No	No phone service	DSL	Yes
Female	No	Yes	No	28	Yes	Yes	Fiber optic	No
Male	No	No	Yes	62	Yes	No	DSL	Yes

***Слика 6**: Преглед података*

StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	No	One year	No	Mailed check	56.95	1889.50	No
No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes
Yes	Yes	Month-to-month	Yes	Electronic check	99.65	820.50	Yes
Yes	No	Month-to-month	Yes	Credit card (automatic)	89.10	1949.40	No
No	No	Month-to-month	No	Mailed check	29.75	301.90	No
Yes	Yes	Month-to-month	Yes	Electronic check	104.80	3046.05	Yes

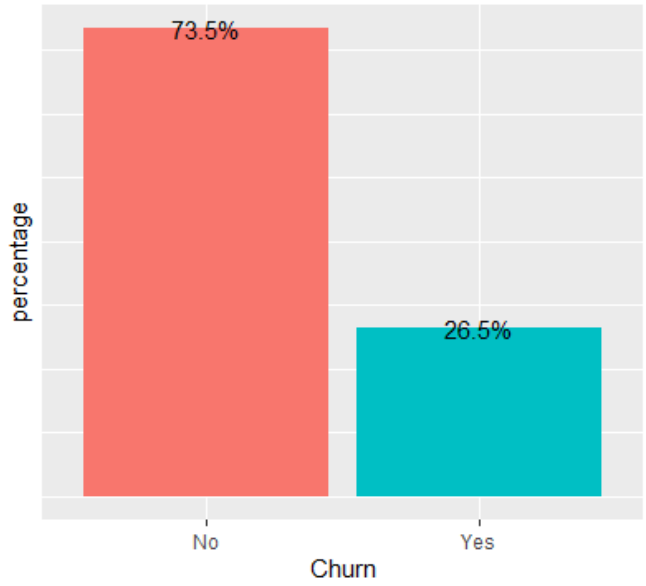
***Слика 7**: Преглед података, наставак*

На почетку рада, скуп података обухватао је 7043 опсервације(корисника) описаних помоћу 21. атрибута(променљиве) приказаних на сликама 6. и 7. Када је реч о типовима података, променљиве које описују кориснички рачун, односно месечна наплата и укупна наплата као и променљива која указује на лојаланост корисника(*tenure*), јесу нумеричког типа док су све остале категоричке, што се јасно може видети из прегледа података.

3.2. Визуелизација података

У наставку дат је графички приказ променљивих(нумеричких и категоричких) на основу којих је касније извршен одабир атрибута који су од значаја за даљу анализу.

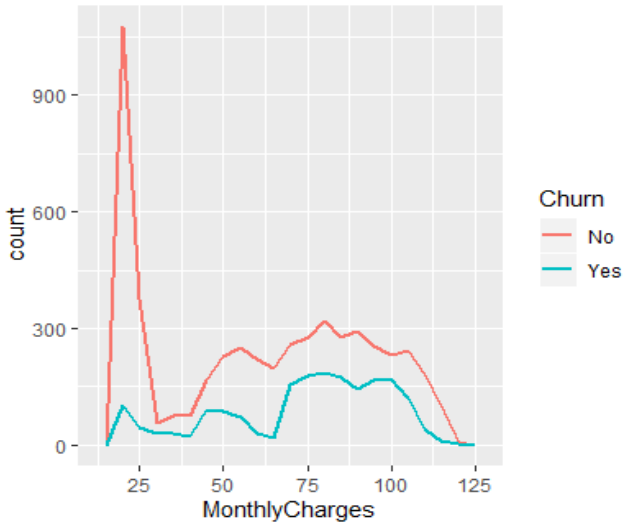
На самом почетку укупан одлазак клијената, на основу историјских података, приказан је на следећем графикону:



***Графикон 1**: Процентуални однос клијената у односу на променљиву churn*

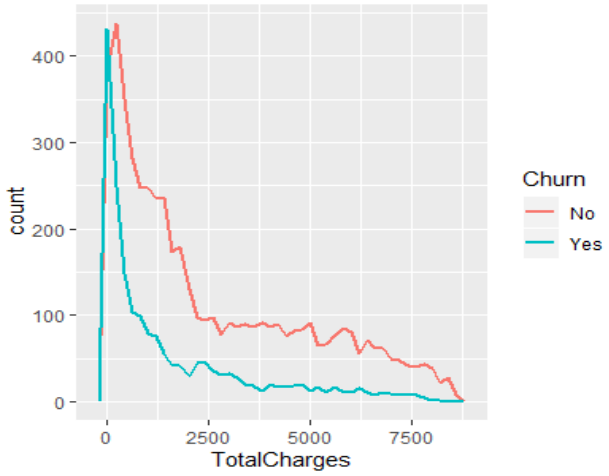
Скоро једна четвртина из датог скупа корисника, одлази односно напушта компанију, што представља не мали број људи.

Графички приказ нумеричких променљивих, месечна и укупна накнада:



***Графикон 2**: Графички приказ променљиве MonthlyCharges у односу на churn*

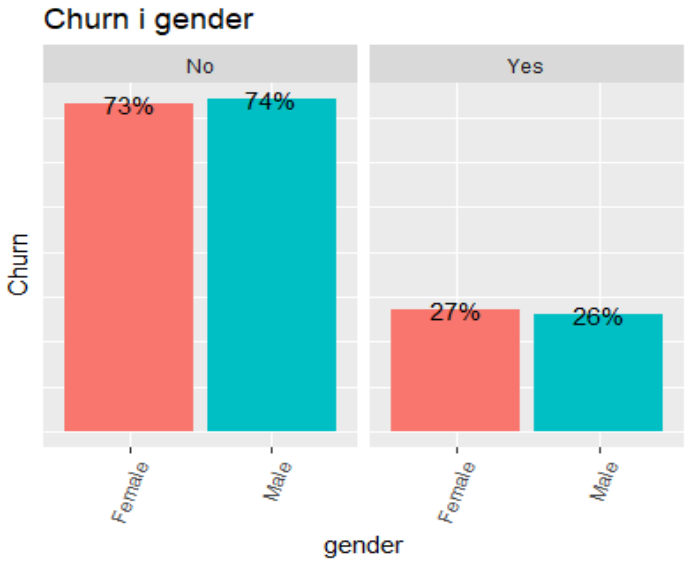
Број оних корисника који нису напустили и при чему плаћају месечно испод 25$ је изузетно велики. Дистрибуција корисника који месечно плаћају преко 30$ делимично је изједначена у односу на *churn.* Такође, корисници који плаћају око 70$ месечно имају највећу тенденцију напуштања компаније.



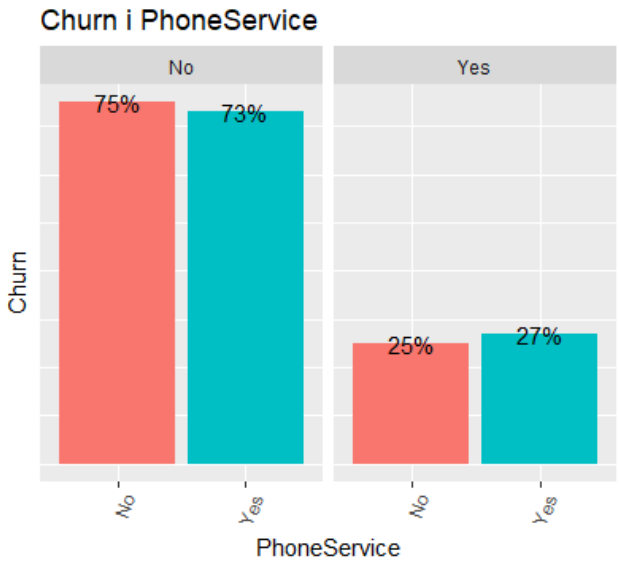
***Графикон 3**: Графички приказ променљиве TotalCharges у односу на churn*

Са графикона се уочава изузетна позитивна асиметричност(дугачак реп расподеле) када је реч о променљивој *TotalCharges*, без обзира да ли је реч о текућим корисницима или о онима који су напустили компанију. Компанију више напуштају они корисници који имају мању вредност ове променљиве, односно они који нису дуго привржени једној компанији.

У даљем наставку, дат је приказ категоричких променљивих у односу на *churn*:

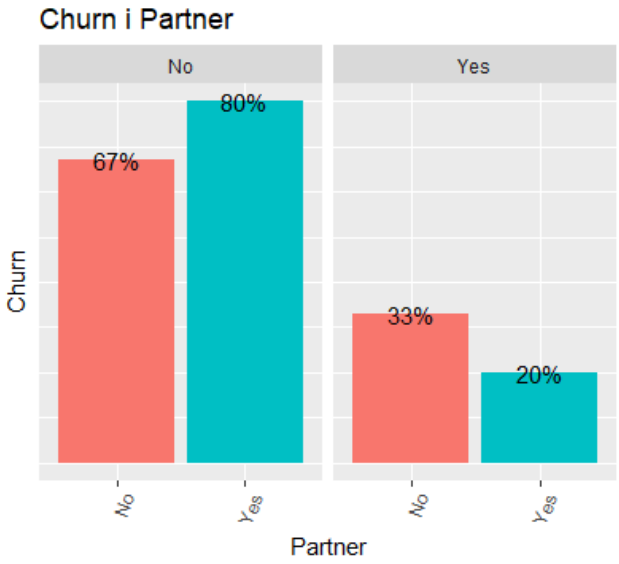


***Графикон 4**: Пол у односу на churn*

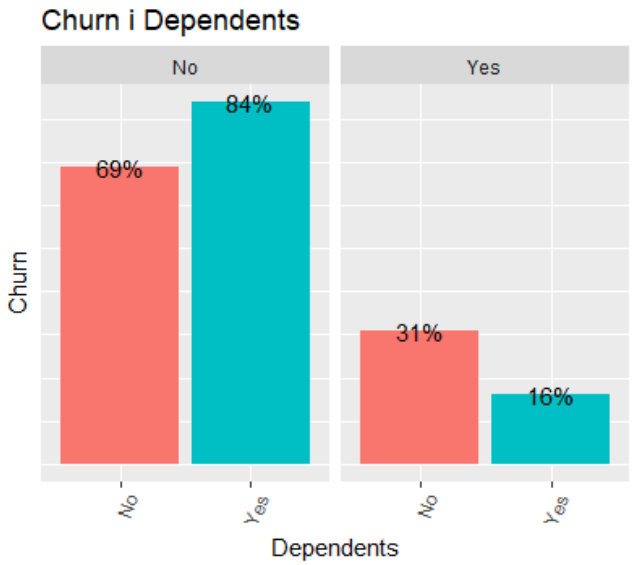


***Графикон 5**: PhoneService у односу на churn*

Са графикона 4., уочавамо да нема неке разлике у томе да ли компанију више напушта мушки или женски пол тј. уједначени су. Исти случај је и са телефонском мрежом(атрибут *PhoneService*).

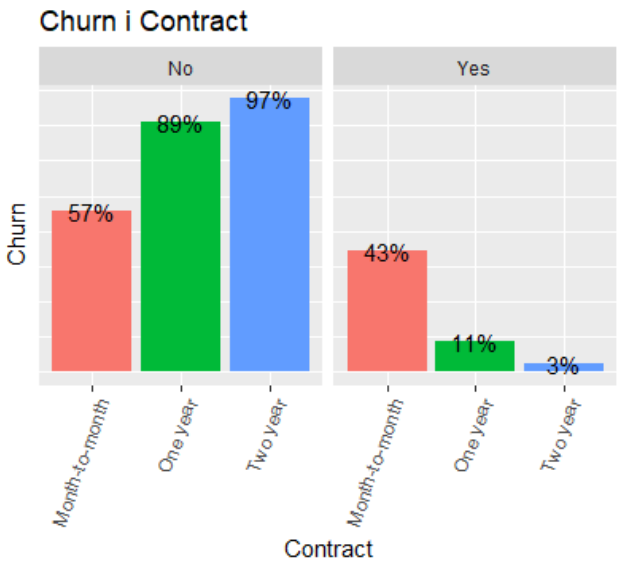


***Графикон 6**: Корисници који имају партнера у односу на churn*



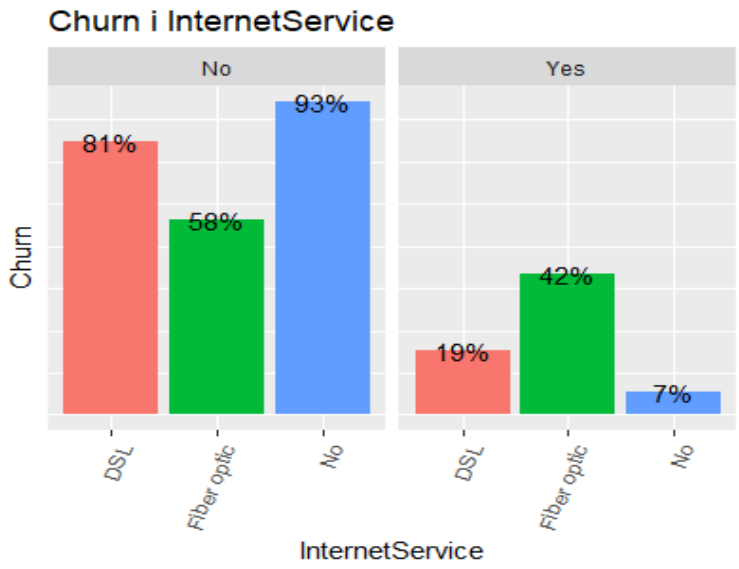
***Графикон 7**: Корисници са децом у односу на churn*

Графикони 6. и 7. приказују понашање корисника у зависности од тога да ли имају партнера или децу. Од свих оних који су прекинули коришћење услуге чак трећина њих (33%) нема партнера и нешто мање(31%) немају децу. Па можемо закључити да они који имају децу и партнера немају толико изражену тенденцију прекидања коришћења услуге.



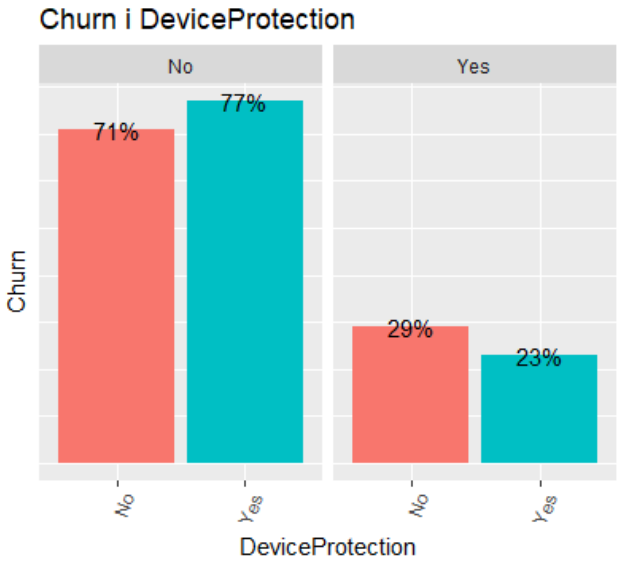
***Графикон 8**: Тип уговора у односу на churn*

Графикон изнад врло илустративно показује да скоро половина, чак 43%, напушта компанију а да притом има уговор који је базиран на месечном нивоу. Док корисници који имају уговоре на две године скоро уопште не доводе у питање раскид уговора, па се може закључити да имају поверења и да су задовољни услугом коју добијају.

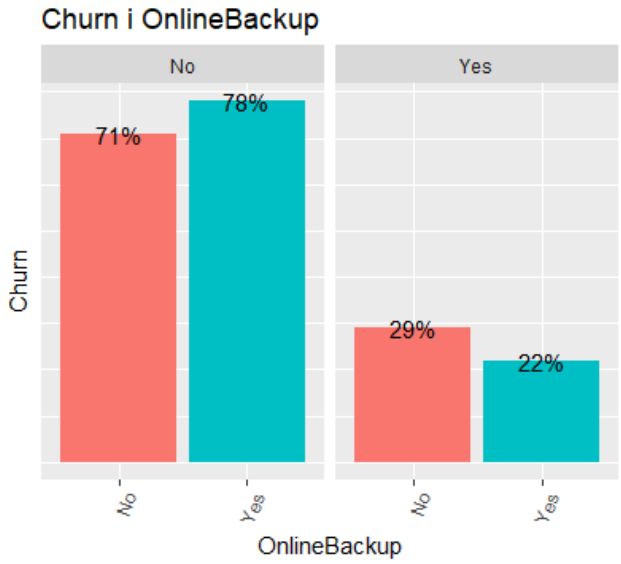


***Графикон 9**: Интернет услуга у односу на churn*

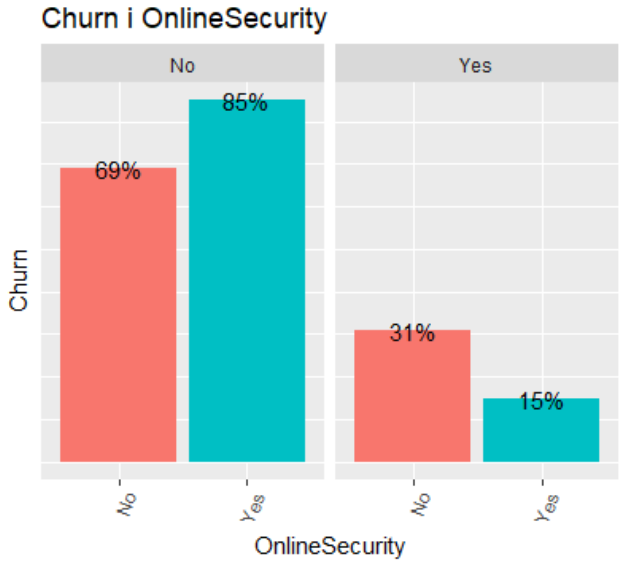
Графикон 9. јасно показује да највећи број корисника одлази а да притом имају интернет спроведен преко оптичког кабла.



***Графикон 10**: DeviceProtection у односу на churn*

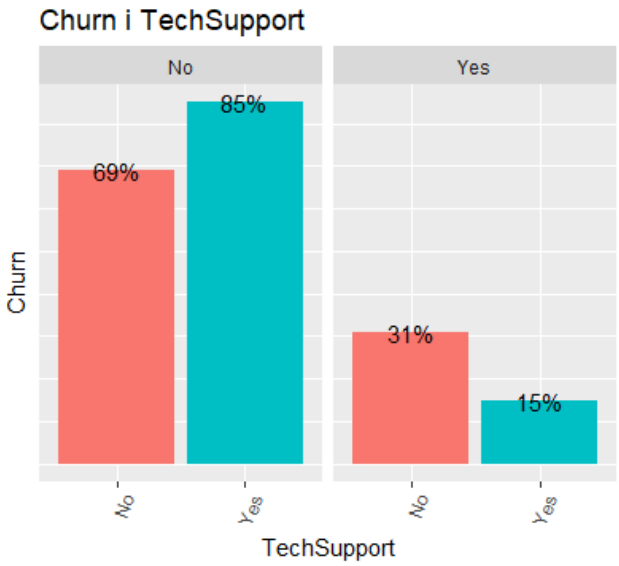


Графикон 11: *Могућност повратка података у односу на churn*



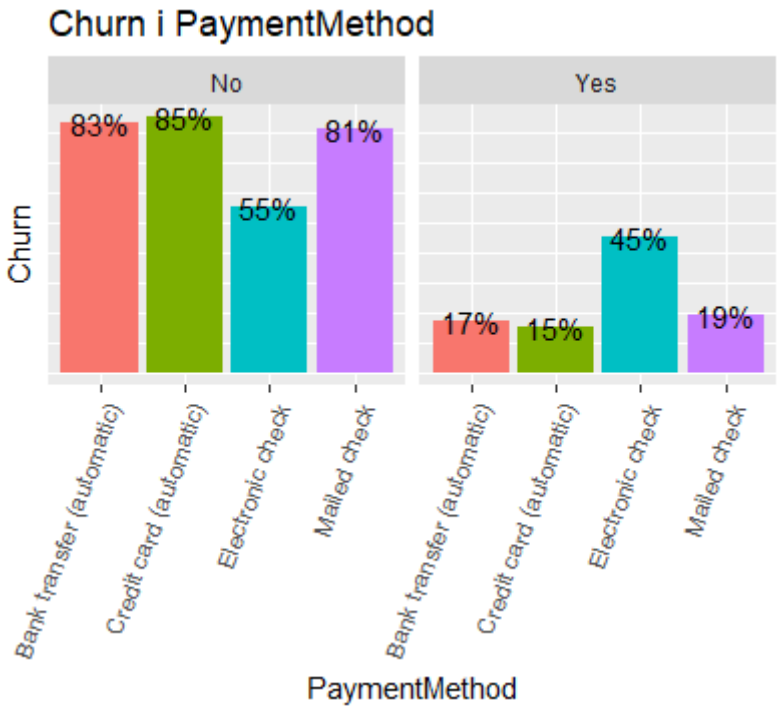
Графикон 12*: Онлине заштита у односу на churn*

Графикони 10., 11. и 12. показују да су, корисници који немају могућност онлајн заштите на интернету, могућност повратка података као и заштиту самог уређаја, више склони одласку у односу на оне који имају претходно наведене могућности. Разлог за то је да неке додатне функционалности корисници знатно више цене и самим тим су задовољнији усугом коју добијају.



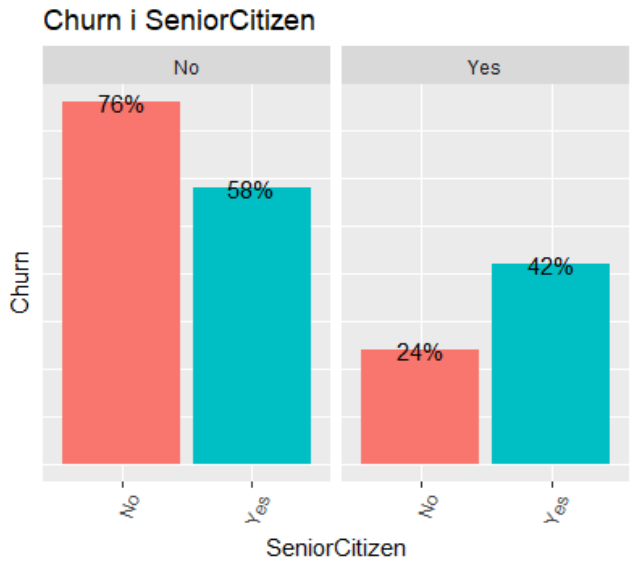
***Графикон 13**: Техничка подршка у односу на churn*

Очигледно је да су, према графикону 13., корисници који имају техничку подршку у склопу свог пакета, лојалнији и да не мењају телекомуникациону компанију.



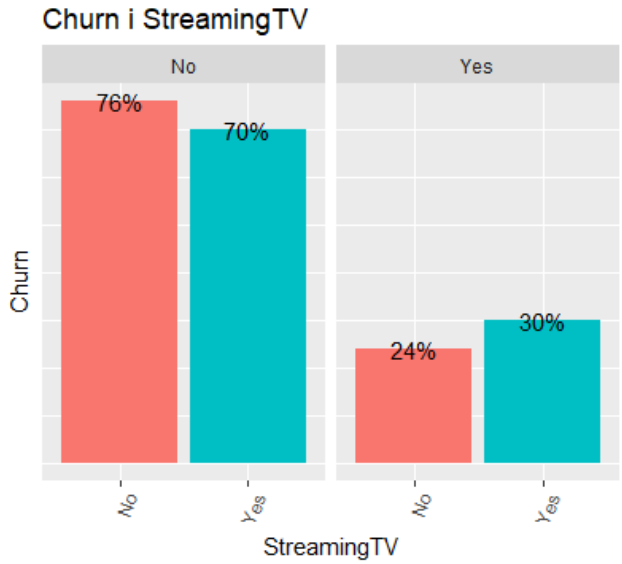
***Графикон 14:** Начин плаћања у односу на churn*

Када је реч о начину плаћања, највећи број корисника који одлазе, јесу заправо они људи који своје рачуне плаћају преко електронских чекова. Док највећи број задовољних корисника јесу они који своје рачуне измирују аутоматски(путем кредитних картица или банкарским трансакцијама).

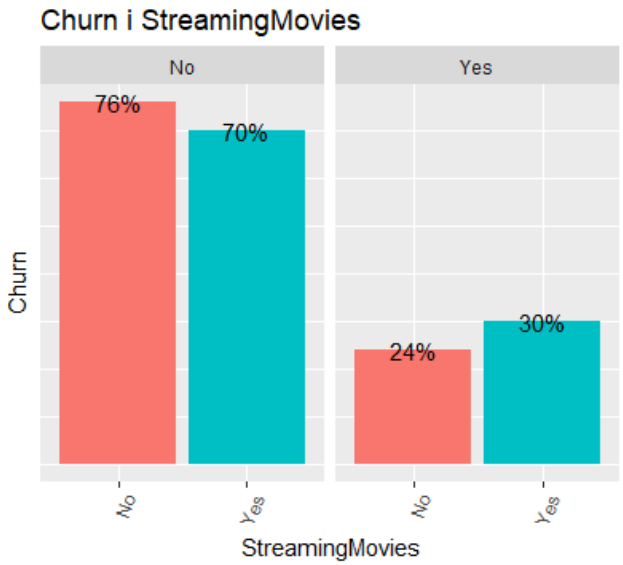


***Графикон 15**: Старост корисника у односу на churn*

Са графикона 15., може се јасно видети да највећи број одлазака предузимају корисници који су старији, при чему је старосна граница дефинисана интерно, у оквиру компаније.

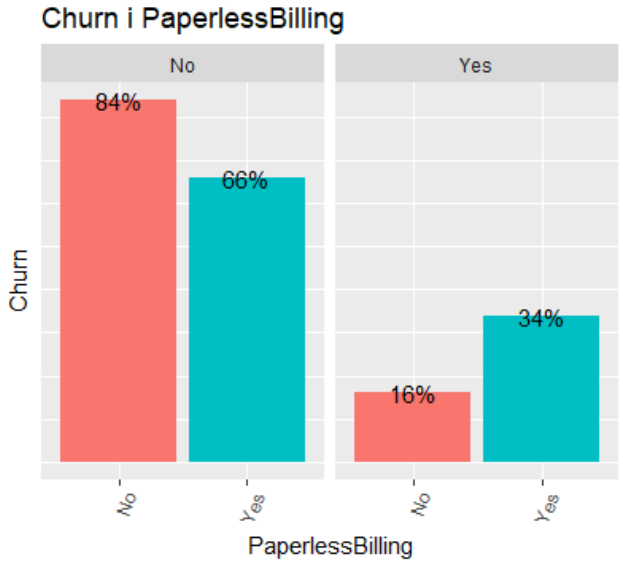


***Графикон 16**: Могућност ТВ стриминга у односу на churn*



***Графикон 17**: Могућност стримовања филмова у односу на churn*

Када је реч о могућности стримовања, ипак већи број корисника одлази иако има ту могућност, па се може закључити да та могућност није пресудна у одлуци коју корисници доносе. Такође, слична ситуација се јавља када су у питању атрибути који указују на могућност коришћења више телефонских линија.

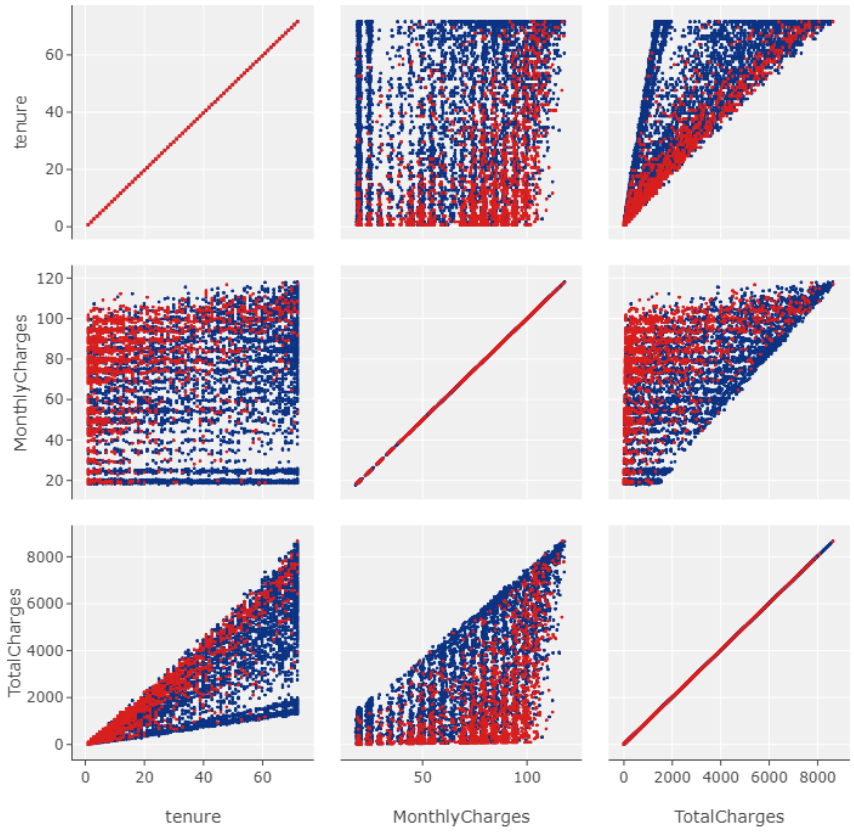


***Графикон 18**: Наплата без папира у односу на churn*

Са графикона 18., јасно се уочава да постоји разлика између корисника који своје рачуне плаћају у папирном облику од оних који то раде онлине или на неки други начин.

3.3. Анализа корелације

На графикону испод, приказан је дијаграм расипања(енгл. *scatterplot matrix*), на коме се могу извести претпоставке о линеарној вези. Посматрају се нумеричке променљиве(месечна и укупна плаћања као и укупан број месеци које је корисник провео у компанији) и њихова зависност. На дијаграму се може уочити јака линеарна веза између месечних накнада и укупних накнада, са израчунатим Pearson-овим коефицијентом корелације од 0.651, док је веза зависности између укупних накнада и броја месеци које је корисник провео као корисник компаније изузетно јака и износи 0.826. Pearson-ов коефицијент корелације се користи у оним случајевима када између променљивих посматраног модела постоји линеарна зависност. Вредност Pearson-овог коефицијента се креће од -1(савршено негативна корелација) до +1(савршено позитивна корелација) при чему предзнак одређује смер корелације.



***Графикон 19**: Дијаграм расипања за нумеричке променљиве*

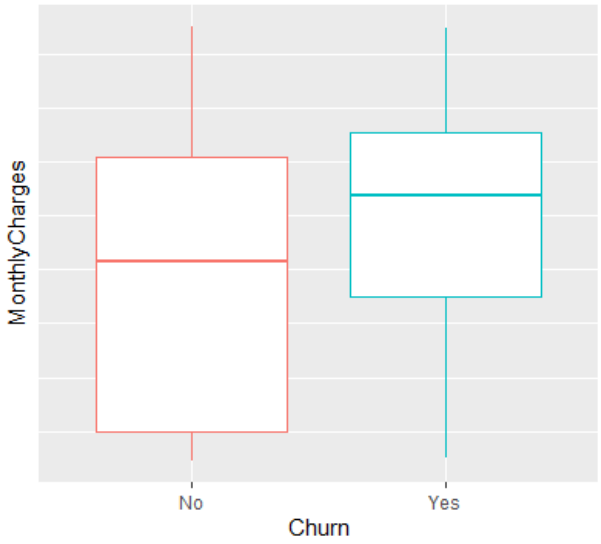
4. Претпроцесирање података

4.1. Припрема података

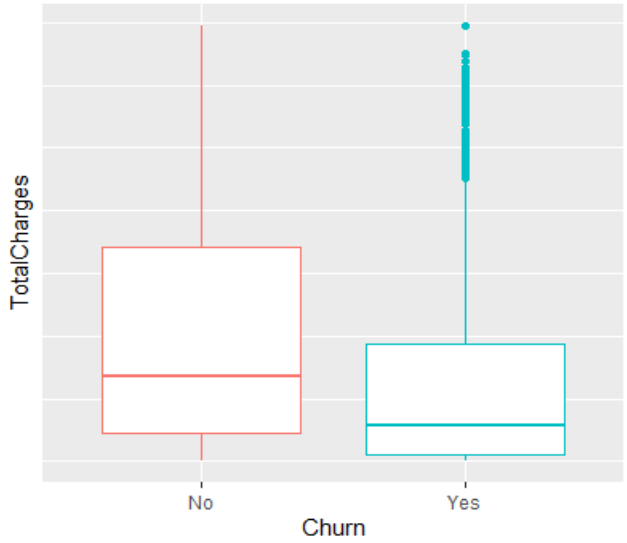
Чишћење података(енгл. *data cleaning*) је поступак припреме података за анализу уклањањем или модификацијом података који су нетачни, непотпуни, небитни, дуплирани или неправилно обликовани. Ови подаци обично нису корисни када је у питању анализа података, јер могу ометати процес или дати нетачне резултате. Постоји неколико метода за чишћење података које зависе од тога како се они посматрају(Han, J., Pei, J., and Kamber, M., 2011).

Конкретно, у раду постоје свега 11 недостајућих вредности и то за променљиву која описује укупна плаћања(*TotalCharges*). Анализом је утврђено да се ради о новим клијентима који су тек потписали уговор, па је самим тим број месеци који су провели у компанији(*tenure*), нула. Такви клијенти су искључени из даље анализе.

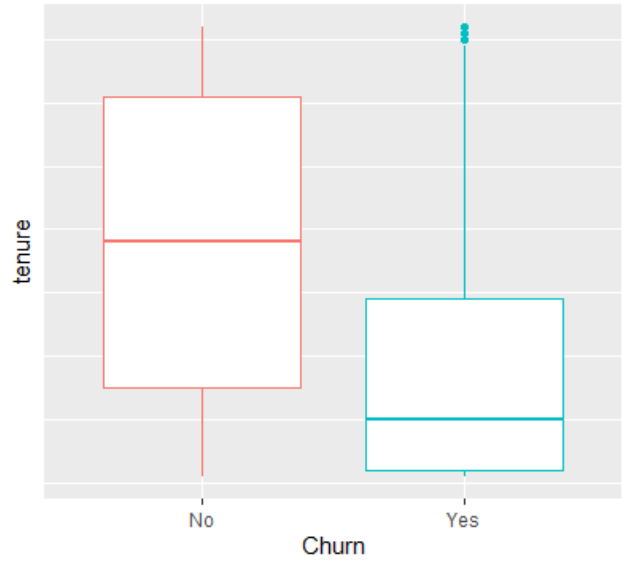
Екстремне вредности(енгл. *outliers*) нумеричких променљивих представљене су на следећим графиконима:



***Графикон 20**: Екстемне вредности променљиве MonthlyCharges*



***Графикон 21**: Екстремне вредности променљиве TotalCharges*



***Графикон 22**: Екстремне вредности променљиве tenure*

Са графикона се јасно уочава да променљива *MonthlyCharges* нема екстремне вредности, док променљива *TotalCharges* поседује велики број таквих вредности. Међутим, променљива *TotalCharges* је касније избачена из анализе(разлози су дати у наставку) па такве вредности нису ни биле предмет даље обраде. Променљива *tenure* поседује три екстремне вредности. То су они корисници који су дуго присутни у самој компанији. Такве вредности нису сређиване јер је сматрано да је скуп података довољно велики(7032) тако да оне не утичу на резултате који су добијени.

Атрибут *SeniorCitizen* поседује вредности 0 и 1, тако да је дати атрибут факторисан.

Такође, атрибути *OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV* и *StreamingMovies* захтевају интернет конекцију, а атрибут *MultipleLines* захтева

телефонску услугу. На основу тога, вредности атрибута као што су “*No internet service*” и “*No phone service*” замењене су само са “*No*”, при чему су остали категорички.

4.2. Одабир атрибута

Избор атрибута извршен је како би се у разматрање узели само они који потенцијално могу бити добри предиктори у моделу. Атрибут *customerID* је искључен јер сам по себи није релевантан за анализу. Атрибути *MonthlyCharges* и *TotalCharges* су међусобно високо корелисани, односно са порастом једне расте и друга, тако да је у разматрање узета само једна променљива, а то је *MonthlyCharges*.

Атрибути *gender*, *phoneService* и *MultipleLines* су искључени из даље анализе јер се нису показали као предиктори који би били потенцијално значајни.

Наравно, постоји много различитих метода за аутоматизацију процеса селекције атрибута. Једна група техника је селекција атрибута похлепним претраживањем подскупова(енгл. *Greedy search*). Друге методе укључују анализу главних компоненти(енгл. *Principal components analysis*) и факторску анализу које се могу користити за редуковање димензија скупа података пројектовањем података у просторе нижих димензија. Како сврха овог рада није претпроцесирање података у наставку се није разматрано о овим темама.

Скуп података са којим су рађене даље анализе укључује 7032 опсервације и 16 атрибута.

5. Вредновање резултата

У овом поглављу су, на почетку, укратко објашњене метрике које су коришћене за упоредну анализу модела описаних у претходном поглављу. Након тога су анализирани конкретни резултати добијени за сваки од модела.

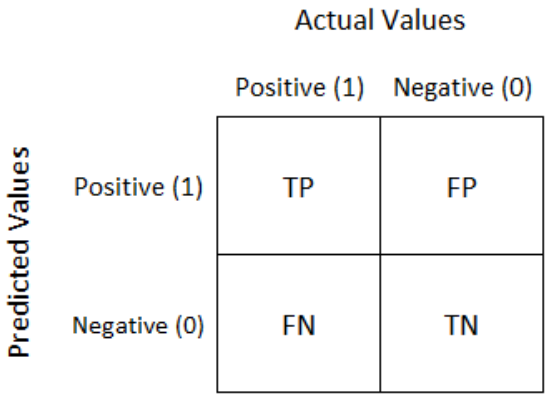
5.1. Тренинг и тест скуп података

Приликом конструисања модела класификације потребно је издвојити део скупа инстанци за тестирање. Тест подаци се не користе у фази креирања модела, како се не би увела пристрасност и добили резултати који су превише оптимистични. Постоји неколико различитих начина да се одаберу скуп података за тренинг и тестирање.

Једна од опција је да се случајно одаберу отприлике једна трећина података(70/30%) за тестирање модела. Такав начин поделе је и извршен у овом раду, односно 70% података одвојено је за тренинг а 30% за тест. Ово је познато као метода раздвајања почетног скупа података, а евалуационе метрике перформанси се израчунавају на основу учинка класификатора на податке који су издвојени за тестирање.

5.2. Евалуационе метрике

Код проблема бинарне класификације, као што је проблем обрађен у овом раду, за сумирање добијених резултата предикције и њихово лакше разумевање, користи се матрица конфузије(енгл. *Confusion matrix*). Ова матрица се састоји од укрштених стварних и предвиђених вредности за припадност класи, и помоћу ње се могу донети закључци о томе колико модел добро врши предикцију излазне променљиве при чему је за позитивну класу узето "*Yes*".



***Слика 8**: Матрица конфузије*[6]

Матрица конфузије се састоји од следећих вредности:

- ***True positives(TP)*** - број исправно разврстаних опсервација у односу на позитивну класу.
- ***False negatives(FN)*** - број погрешно разврстаних опсервација у односу на негативну класу.
- ***False positives(FP)*** - број погрешно разврстаних опсервација у односу позитивну класу.
- ***True negatives(TN)*** – број исправно разврстаних опсервација у односу на негативну класу.

Добијене вредности у матрици конфузије се даље користе за израчунавање основних евалуационих метрика које представљају показатеље успешности предвиђања тестираног модела. Основне метрике се добијају помоћу следећих формула и у њих спадају:

- ***Accuracy(тачност)*** је метрика која показује колики је проценат опсервација правилно класификован у класу којој заиста и припада.

$$\frac{TP + TN}{TP + TN + FP + FN} \qquad 5.1.$$

[6] https://towardsdatascience.com/taking-the-confusion-out-of-confusion-matrices-c1ce054b3d3e

- ***Precision(прецизност)*** је метрика која показује вероватноћу да је предвиђена позитивна класа, а да је заиста и била позитивна.

$$\frac{TP}{TP + FP} \qquad 5.2.$$

- ***Recall(одзив)*** је метрика која показује вероватноћу да се десила позитивна класа, а да је заиста и била предвиђена позитивна класа.

$$\frac{TP}{TP + FN} \qquad 5.3.$$

- ***F1*** метрика се дефинише као хармонијска средина између две величине(Han, J., Pei, J., and Kamber, M., 2011).

$$F_1 - score = \frac{2 * precision * recall}{precision + recall} \qquad 5.4.$$

6. Анализа резултата

У следећој табели наведене су све вредности евалуационих метрика за сваки модел објашњен у поглављу 2.

***Табела 1**: Поређење евалуационих метрика свих креираних модела*

	Accuracy	Precision	Recall	F1
glm.fit.1	0.7945920	0.5357143	0.6342495	0.5808325
glm.fit.2	0.7879507	0.6357143	0.5943239	0.6143227
glm.fit.3	0.7623340	0.7482143	0.5378691	0.6258402
glm.fit.4	0.7025617	0.8553571	0.4673171	0.6044164
lda	0.7974383	0.5589286	0.6348884	0.5944919
qda	0.7580645	0.7392857	0.5321337	0.6188341
knn	0.7794118	0.4178571	0.6273458	0.5016077
svm.1	0.7727704	0.4500000	0.5957447	0.5127162

Модели чије су евалуационе метрике представљене у Табели 1. су:

- **Glm.fit.1** – модел логистичке регресије са подразумеваном вредношћу параметра *treshhold*(0.5).
- **Glm.fit.2** – модел логистичке регресије са оптимизованом вредношћу параметра *treshhold*(0.4).
- **Glm.fit.3** – модел логистичке регресије са оптимизованом вредношћу параметра *treshhold*(0.3).
- **Glm.fit.4** – модел логистичке регресије са оптимизованом вредношћу параметра *treshhold*(0.2).
- **Lda** – модел линеарне дискриминационе анализе
- **Qda** – модел квадратне дискриминационе анализе
- **Knn** – модел К најближих суседа
- **Svm.1** – модел методе потпорних вектора

На основу Табеле 1., може се закључити, судећи по евалуационим метрикама, да сви креирани модели дају добре резултате предвиђања. Метрика тачности се за све моделе креће приближно једнако(приближно вредности 0.8) са изузетком модела логистичке регресије са вредношћу параметра *treshhold* 0.2 која је нешто нижа и износи 0.702. Прецизност је дефинасана у опсегу од 0.41 до 0.85, што представља прилично широк опсег. Може се закључити да најмању прецизност има модел К најближих суседа(0.41) и нешто мало бољу прецизност има метод потпорних вектора(0.45). Остали модели имају бољу ову метрике док модел логистичке регресије са вредношћу параметра *treshhold* 0.2 иако има најмању тачност, ипак има највећу прецизност, чак 0.85. Када је реч о метрици која описује одзив, углавном све имају приближно једнаке вредности, односно од 0.59 до 0.63, са изузетком модела логистичке регресије са вредностима параметра *treshhold* од 0.3 и 0.2 као и квадратне дискриминационе анализе. *F1* метрика, као можда најважнија, има највећу вредност за модел логистичке регресије са вредношћу параметра *treshhold* од 0.3. Такође, и модел квадратне дискриминационе анализе даје приближно сличне резултате.

Метрика која описује прецизност модела тј. показује однос тачно предвиђених(TP) и све предвиђене унутар класе(TP + FP) даје најбоље резултате за модел *glm.fit.4*. На основу тога компанија може тачније одредити вероватноћу одласка следећег првог корисника. Међутим, како је проблем овог рада идентификација оних клијената који ће вероватно отказати услуге телекомуникационе компаније, ипак ће доносиоцима одлука од највећег значаја бити метрика која описује одзив. Она показује вероватноћу да се десила позитивна класа, а да је заиста и била предвиђена позитивна класа, односно минимизација броја оних клијената који су напустили компанију, а предвиђено је да неће.

7. Закључак

Циљ овог рада био је да се применом машинског учења, конкретно метода класификације, дође до одређених резултата и закључака који би допринели бољем пословању компаније. Посебан допринос овај рад може да пружи запосленима у компанији да разумеју потребе корисника и да покушају да на основу донетих закључака пронађу механизме за остварење једног од основних циљева компаније, а то је задржати што већи број задовољних корисника.

Након дефиниције самог проблема који је разматран у овом раду, укратко је објашњен процес функционисања сваког модела. Модели су креирани применом алгоритама дискриминационе анализе, логистичке регресије, методе потпорних вектора и методе К најближих суседа. Приликом примене сваког од модела водило се рачуна да се нађу оптимални параметри тог модела како би се добили најбољи резултати.

Пре креирања самих модела, извршена је анализа читавог скупа података. Већина атрибута имала је релативно јасно значење, тако да су уклоњене само оне променљиве, које према процени, не би дале значајан допринос у креирању предиктивних модела.

Након претпроцесирања података, које обухвата припрему и одабир података, дефинисани су параметри вредновања самих резултата. Они су исказани помоћу следећих евалуационих метрика: тачност, прецизност, одзив и Ф-1 скор. Битно је нагласити да не постоји најбољи модел. Компанија мора сама одлучити који од модела је најбољи за њу и који ће у датом тренутку дати најбољи могући напредак у пословању или да се служи комбинацијом више њих. Такође, сами алгоритми показали су да су променљиве које описују лојалност корисника(*tenure*), једногодишњи и двогодишњи уговори као и променљива *PaperlessBilling*, имају највећу значајност. Тако да се може извршити додатно испитивање корисника који имају овакав скуп карактеристика јер ти атрибути највише утичу и имају највећу значајност приликом креирања модела.

При припреми и претпроцесирању података, приступило се одабиру атрибута на интуитиван и рационалан начин који је такође и графички употпуњен. Међутим, треба напоменути да постоје и статистичке методе које врше редукцију димензија скупа података применог математичких модела(факторска анализа, метода главних компоненти).

Након одабира атрибута који ће бити укључени у моделе, извршена је и њихова имплементација. Креирана су четири модела логистичке регресије са различитом вредношћу *treshhold*-а. Први модел логистичке регресије даје најбоље резултате када су у питању тачност(0.79) и одзив(0.63) модела. Највећу прецизност има четврти модел логистичке регресије, 0.79. Највећу вредност метрике која описује хармонијску средину, Ф-1 скор, има трећи модел(0.62) и он је уједно и највећи код свих модела. Линеарна дискриминациона анализа има тачност од 0.79, док је прецизност 0.55 а одзив 0.63. Квадратна дискриминациона анализа има лошије резултате од линеарне осим када су у питању прецизност(0.73) и Ф-1 скор(0.61). Модел К најближих суседа је оптимизован и добијена вредност за К је 17. Тачност овог модела је 0.77, прецизност 0.41 а одзив 0.62. Метод потпорних вектора има тачност од 77%, прецизност 45% док је одзив 59%. Најбољи резултати зависе од модела до модела.

Спроведене анализе могу се евентуално додатно унапредити коришћењем неких напреднијих алгоритама. Један од таквих алгоритама јесте *Extreme Gradient Boosting* алгоритам. Овај алгоритам је заснован на принципима *gradient boosting framework*-а, који је један од популарнијих алгоритама међу корисницима *R* програмског језика.

На крају, важно је напоменути, да нема никаквих гаранција да ће резултати бити бољи зато што су коришћени напреднији алгоритми, јер и други фактори утичу на резултате. Може се дести да алгоритми који су се показали као изузетно добри приликом анализе једног скупа података могу имати значајно слабије резултате при анализи другог скупа података. Зато је веома важно разумети скуп података и његове атрибуте на основу којих се врше предикције.

8. Литература

Anton, Howard. (1994). *Elementary Linear Algebra.*

Cortes Corinna, Vapnik, Vladimir N. (1995). *Support-vector networks.* CiteSeerX.

Gareth M. James, Trevor J. Hastie. (2002). *Functional linear discriminant analysis for irregularly sampled curves.* Los Angeles: Royal Statistical Society.

Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques.* The Morgan Kaufmann Series in Data Management Systems. Elsevier Science.

Hsu, Chih-Wei, Chang, Chih-Chung & Lin, Chih-Jen. (2003). *A Practical Guide to Support Vector Classification.* Department of Computer Science and Information Engineering, National Taiwan University.

Inc., Sas Institute. (2018). *Machine Learning - What it is and why it matters.* Преузето са: https://www.sas.com/en_us/insights/analytics/machine-learning.html#machine-learning-users.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R.* (S. T. Statistics, Ур.) New York: Springer.

Onel Harrison. (2018). *Machine Learning Basics with the K-Nearest Neighbors Algorithm.* Преузето са: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

Rouse. (2018). *Special Report: Artificial intelligence apps come of age.* Преузето са: https://searchenterpriseai.techtarget.com/definition/machine-learning-ML.

Stacker, M. (2015, April). Преузето са https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/

Walker SH, Duncan DB. (1967). *Estimation of the probability of an event as a function of several independent variables.* Biometrika.

9. Прилог

У овом делу дата је имплементација у програмском језику *R*, при чему је коришћено развојно окружење *Rstudio*.

Учитавање, припрема података и избор атрибута:

```
# Avgust, 2019
# Telecom customer churn

# Instaliranje i ucitavanje potrebnih biblioteka
install.packages("tidyverse")
install.packages("cowplot")
install.packages("caret")
install.packages("rpart")
install.packages("ROCR")
install.packages("rpart.plot")
library(tidyverse)
library(cowplot)
library(caret)
library(rpart)
library(ROCR)
library(rpart.plot)
install.packages("ggplot")
library(ggplot2)

# Ucitavanje dataset-a
churn <- read.csv(file = "telecom customer churn.csv")

# Pregled dataset-a kao i tipova varijabli
summary(churn)
str(churn)

# Varijabla TotalCharges sadrzi 11 NA vrednosti koje je potrebno otkloniti
churn[is.na(churn$TotalCharges),1:6]
churn <- churn[complete.cases(churn),]
dim(churn)

# Varijablu SeniorCitizen je potrebno faktorisati buduci da ima vrednosti 0 ili 1
summary(churn$SeniorCitizen)
str(churn$SeniorCitizen)
unique(churn$SeniorCitizen)
churn$SeniorCitizen <-
    as.factor(ifelse(test = churn$SeniorCitizen == 1, yes = "Yes", no = "No"))
str(churn)
table(churn$SeniorCitizen)

# Cuvanje trenutnog dataset-a, kao prve, inicijalne verzije skupa podataka
saveRDS(object = churn, file = "telecom customer churn, v1.1.RData")

# Priprema podataka za modele klasifikacije
## Monthly charges i total charges su visoko korelisani tako da se jedna izbacuje (Total
Charges)
## Uklanjanje customerID koja nije relevantna za analiziranje

churn$customerID <- NULL
churn <- churn[,-c(19)]

```

```
50. ##########################################################################################
    ########
51.
52. ## ggplot theme
53. theme <- theme(
54.   axis.text.y = element_blank(), axis.ticks.y = element_blank(),
55.   legend.position="none"
56. )
57.
58. summary(churn)
59. str(churn)
60. saveRDS(object = churn, "telecom customer churn, selected features, v1.1.RData")
```

Визуелизација података:

```
1.  # Vizuelizacija podataka
2.
3.  attach(churn)
4.  summary(churn)
5.  data <- read.csv("telecom customer churn.csv")
6.
7.  # Najpre, analiziramo numericke varijable i njihove raspodele:
8.  library(ggplot)
9.  ggplot(data = churn, aes(MonthlyCharges, color = Churn))+
10.   geom_freqpoly(binwidth = 5, size = 1)
11. # Broj tekucih korisnika ciji mesecni racun iznosi do 25$ je znacajno veliki,
12. # dok je distribucija korisnika sa mesecnim racunom preko 30$ delimicno izjednacena u od
    nosu na churn
13.
14. # MonthlyCharges u odnosu na churn
15. t.test(MonthlyCharges ~ Churn, var.equal = T)
16. wilcox.test(MonthlyCharges~Churn)
17. summary(churn)
18. ?t.test
19.
20. ggplot(data = data, aes(TotalCharges, color = Churn))+
21.   geom_freqpoly(binwidth = 200, size = 1)
22. # Sa grafika uocavamo izuzetnu pozitivnu asimetricnost (dugacak rep raspodele) kada
23. # je rec o varijabli TotalCharges, bez obzira da li se radi o tekucim korisnicima ili
24. # onima koj su napustili kompaniju.
25.
26. ggplot(data = data, aes(tenure, color = Churn))+
27.   geom_freqpoly(binwidth = 200, size = 1)
28. # Tenure u odnosu na churn
29. churn$tenure <- as.numeric(churn$tenure)
30. t.test(tenure ~ Churn, var.equal = T)
31. levels(churn$Churn)
32. summary(churn)
33. ?t.test
34.
35. # Korelacija izmedju tenure, MC i TC, dijagram rasipanja
36. ?cor
37. data$tenure <- as.numeric(data$tenure)
38. str(data)
39. corrplot::corrplot.mixed(cor(cor))
40. cor <- as.matrix(data[,c(6,19,20)])
41. cor
42.
43. options(repr.plot.width = 4, repr.plot.height = 3)
44. library(tidyverse)
```

```
# Procentualni prikaz korisnika koji su napustili kompaniju (varijabla churn)
data %>%
  group_by(Churn) %>%
  summarize(n = n()) %>%
  mutate(
    percentage = round(n / sum(n), 3),
    n = NULL) %>%
  ggplot(aes(x = Churn, y = percentage)) + geom_col(aes(fill = Churn)) +
  theme +
  geom_text(
    aes(x = Churn, y = percentage, label = paste(percentage*100, "%", sep = ""))
  )
# 26.5% korisnika u ovom skupu podataka je napustilo kompaniju

##############################################################################################
###########

# Smanjivanje velicine grafika
options(repr.plot.width = 4, repr.plot.height = 4)

# Prikaz varijable churn u odnosu na kategoricke varijable, graficki
## Kategoricke (faktorske) varijable za prikaz
function_columns <- churn %>%
  select(
    "gender", "SeniorCitizen", "Partner", "Dependents", "PhoneService", "MultipleLines",
    "InternetService", "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport",
    "StreamingTV", "StreamingMovies", "Contract", "PaperlessBilling", "PaymentMethod","Churn"
  )

for (i in 1:ncol(function_columns)) {
  # Get column names so dplyr group by works
  cname <- colnames(function_columns[c(i,17)])
  # Subset data frame by variable name selected
  a <- subset(
    function_columns, !is.na(function_columns[,i]) & function_columns[,i] != "",
    select = cname
  ) %>%
    # Create percentage statistics per variable
    group_by_at(vars(cname)) %>%
    summarize(
      n = n()
    ) %>%
    mutate(
      Percentage = round(n / sum(n), 2)
    )

  # Save plot in a variable so plots can be displayed sequentialy
  p <- ggplot(
    data = a, aes_string(
      x = colnames(a[1]), y = colnames(a[4]), fill = colnames(a[1])
    )
  ) +
    # Split each graph per Churn to see influence of variable
    facet_wrap("Churn") +
    geom_bar(stat = "identity") +
    # Make graph a bit cleaner
          theme(
            axis.text.y = element_blank(), axis.ticks.y = element_blank(),
```

```
            axis.text.x = element_text(angle = 70, hjust = 1),
            legend.position="none"
          ) +
          geom_text(
            aes(y = Percentage, label = paste0(Percentage * 100,"%"))
          ) +
          labs(
            x = colnames(a[1]), y = "Churn", title = paste("Churn i", colnames(a[1]))

          )

        # Display graphs
        print(p)
        # Cleanup
        rm(cname)
        rm(i)
      }

      ################################################################################
  ########################

      options(repr.plot.width = 7, repr.plot.height = 3)
      data %>%
          filter(Churn == "Yes") %>%
          group_by(tenure) %>%
          summarize(
            n = n()
          ) %>%
          mutate(
            Percentage = round(n / sum(n), 3)
          ) %>%
          # Create plot
          ggplot(
            aes(x = tenure, y = Percentage, color = tenure)
          ) +
          stat_smooth(method = "lm", col = "red") +
          geom_point(alpha = 2/3) +
          # Clean graph visual a bit
          theme +
          labs(
            x = "Tenure", y = "Churn (%)"
          )

      ggplot(data = churn, aes(y = tenure, x = Churn, color = Churn)) +
        theme +
        geom_boxplot()
      # Prikaz outlier-
  a za tenure varijablu odnosu na churn. Mogu se uociti 3 outlier-a kada je churn yes.

      data %>%
        filter(Churn == "Yes") %>%
        group_by(MonthlyCharges) %>%
        summarize(
          n = n()
        ) %>%
        mutate(
          Percentage = round(n / sum(n), 3)
        ) %>%
        # Create plot
        ggplot(
          aes(x = MonthlyCharges, y = Percentage, color = MonthlyCharges)
```

```
        ) +
        stat_smooth(method = "lm", col = "red") +
        geom_point(alpha = 2/3) +
        # Clean graph visual a bit
        theme +
        labs(
          x = "MonthlyCharges", y = "Churn (%)"
        )

      ggplot(data = churn, aes(y = MonthlyCharges, x = Churn, color = Churn)) +
        theme +
        geom_boxplot()
      # MonthlyCharges nema outlier-a

      data %>%
        filter(Churn == "Yes") %>%
        group_by(TotalCharges) %>%
        summarize(
          n = n()
        ) %>%
        mutate(
          Percentage = round(n / sum(n), 3)
        ) %>%
        # Create plot
        ggplot(
          aes(x = TotalCharges, y = Percentage, color = TotalCharges)
        ) +
        stat_smooth(method = "lm", col = "red") +
        geom_point(alpha = 2/3) +
        # Clean graph visual a bit
        theme +
        labs(
          x = "TotalCharges", y = "Churn (%)"
        )

      ggplot(data = data, aes(y = TotalCharges, x = Churn, color = Churn)) +
        theme +
        geom_boxplot()
      # Kada je churn yes, TotalCharges ima prilicno veliki broj outlier-
    a. Obratiti paznju.
      str(churn)
```

Коришћене евалуационе метрике:

```
getEvaluationMetrics <- function(cm) {

  TP <- cm[2, 2] # true positive
  TN <- cm[1, 1] # true negative
  FP <- cm[1, 2] # false positive
  FN <- cm[2, 1] # false negative

  accuracy = sum(diag(cm)) / sum(cm)
  precision <- TP / (TP + FP)
  recall <- TP / (TP + FN)
  F1 <- (2 * precision * recall) / (precision + recall)

  c(Accuracy = accuracy,
    Precision = precision,
    Recall = recall,
    F1 = F1)
```

```
}
```

Имплементација модела:

```
# Avgust, 2019
# Telecom customer churn

# Varijable OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV and
 StreamingMovies
# zahtevaju internet konekciju i varijabla MultipleLines zahteva telefonsku uslugu tako
da
# "No internet service" i "No phone service" zamenjujemo sa "No"
factorrenames <- names(churn[9:14])

churn <- churn %>%
  mutate_at(.vars=factorrenames,
            .funs=~recode_factor(., `No internet service`="No")) %>%
  mutate_at(.vars="MultipleLines",
            .funs=~recode_factor(., `No phone service`="No"))

str(churn)

# Podela dataset-a na train i test
library(caret)
set.seed(1)
i <- createDataPartition(churn$Churn, p = 0.7, list = FALSE)
train <- churn[i, ]
test <- churn[-i, ]
dim(train)
dim(test)

######################### LOGISTICKA REGRESIJA ######################
attach(churn)

# Kreiranje model
glm.fit.1 <- glm(Churn ~ ., data = train, family = binomial)
summary(glm.fit.1)
contrasts(Churn)
glm.fit.1.prob <- predict(glm.fit.1, newdata = test, type = "response")
head(glm.fit.1.pred)

# Predikcije
glm.fit.1.pred <- rep("No",nrow(test))
glm.fit.1.pred[glm.fit.1.prob > .5] <- "Yes"

# Matrica konfuzije
glm.fit.1.cm <- table(predicted = glm.fit.1.pred, true = test$Churn)
glm.fit.1.cm

# Ukupna greska
mean(glm.fit.1.pred != test$Churn)

# Evaluacione metrike
source("Evaluacione metrike.R")
glm.fit.1.em.1 <- getEvaluationMetrics(glm.fit.1.cm)
glm.fit.1.em.1
summary(test)

# Promena treshhold-a kako bismo gresku modela koja je 20%
glm.fit.1.pred[glm.fit.1.prob > .4] <- "Yes"
```

```
55.
56. # Matrica konfuzije
57. glm.fit.1.cm <- table(predicted = glm.fit.1.pred, true = test$Churn)
58. glm.fit.1.cm
59.
60. # Ukupna greska
61. mean(glm.fit.1.pred != test$Churn)
62.
63. # Evaluacione metrike
64. source("Evaluacione metrike.R")
65. glm.fit.1.em.2 <- getEvaluationMetrics(glm.fit.1.cm)
66. glm.fit.1.em.2
67. summary(test)
68. # Nalazenje najboljeg tresh hold-a
69.
70. glm.fit.1.pred[glm.fit.1.prob > .3] <- "Yes"
71.
72. # Matrica konfuzije
73. glm.fit.1.cm <- table(predicted = glm.fit.1.pred, true = test$Churn)
74. glm.fit.1.cm
75.
76. # Ukupna greska
77. mean(glm.fit.1.pred != test$Churn)
78.
79. # Evaluacione metrike
80. source("Evaluacione metrike.R")
81. glm.fit.1.em.3 <- getEvaluationMetrics(glm.fit.1.cm)
82. glm.fit.1.em.3
83. summary(test)
84.
85. glm.fit.1.pred[glm.fit.1.prob > .2] <- "Yes"
86.
87. # Matrica konfuzije
88. glm.fit.1.cm <- table(predicted = glm.fit.1.pred, true = test$Churn)
89. glm.fit.1.cm
90.
91. # Ukupna greska
92. mean(glm.fit.1.pred != test$Churn)
93.
94. # Evaluacione metrike
95. source("Evaluacione metrike.R")
96. glm.fit.1.em.4 <- getEvaluationMetrics(glm.fit.1.cm)
97. glm.fit.1.em.4
98. summary(test)
99.
100.        # Tabelarni prikaz za razlicite tresh hold-ove
101.        metrics <-
      data.frame(rbind(glm.fit.1.em.1, glm.fit.1.em.2, glm.fit.1.em.3, glm.fit.1.em.4))
102.        rm(k)
103.
104.        # Cross-validation
105.        install.packages("boot")
106.        library(boot)
107.        ?cv.glm
108.        cv.1 <- cv.glm(train, glm.fit.1)
109.        cv.1$delta
110.        # CV dobijamo da je test greska naseg modela 13%, pa mozemo reci da je zadovolja
    vajuce
111.
112.        metrics
113.        summary(glm.fit.1)
```

```
114.
115.        ########################## Linearna diskriminaciona analiza ####################
    ########
116.
117.        library(MASS)
118.        # Kreiranje modela
119.        lda.fit <- lda(Churn ~ ., data = train)
120.        plot(lda.fit)
121.        attach(churn)
122.
123.        # Kreiranje predikcija
124.        lda.pred <- predict(lda.fit, test)
125.        lda.em <-
     getEvaluationMetrics(table(predicted = lda.pred$class, true = test$Churn))
126.
127.        # Kontrolisanjo dodavanje varijabli. Dodajem varijable koje sam logistickom regr
    esijom zakljucio da
128.        # mogu biti znacajne
129.        lda.fit.1 <- lda(Churn ~ Contract + tenure + PaperlessBilling, data = train)
130.        plot(lda.fit.1)
131.        lda.pred.1 <- predict(lda.fit.1, test)
132.        lda.em.1 <-
     getEvaluationMetrics(table(predicted = lda.pred.1$class, true = test$Churn))
133.        lda.em.1
134.        lda.em
135.
136.        ########################## Kvadratna diskriminaciona analiza ###################
    ########
137.
138.        library(MASS)
139.        # Kreiranje modela
140.        qda.fit <- qda(Churn ~ ., data = train)
141.
142.        # Kreiranje predikcija
143.        qda.pred <- predict(qda.fit, test)
144.        qda.em <-
     getEvaluationMetrics(table(predicted = qda.pred$class, true = test$Churn))
145.        qda.em
146.        lda.em
147.
148.        ########################################## KNN ##################################
    ########
149.
150.        train.s <- train
151.        train.s$tenure <- scale(train.s$tenure)
152.        train.s$MonthlyCharges <- scale(train.s$MonthlyCharges)
153.        train.s.y <- train.s$Churn
154.        train.s <- train.s[,-c(1:4,6:17)]
155.        train.s.2 <- train.s
156.        train.s.2$Churn <- train.s.y
157.
158.        test.s <- test
159.        test.s$tenure <- scale(test.s$tenure)
160.        test.s$MonthlyCharges <- scale(test.s$MonthlyCharges)
161.        test.s.y <- test.s$Churn
162.        test.s$Churn <- test.s.y
163.        test.s.2 <- test.s[,-c(1:4,6:17)]
164.        test.s.2 = test.s.2[,-c(3)]
165.        test.s.y
166.        library(class)
167.
```

```
168.        # problem overfitting-a
169.        knn.pred.1 <- knn(train = train.s, test.s.2, cl = train.s.y, k=1)
170.        getEvaluationMetrics(table(knn.pred.1, test.s.y))
171.
172.        knn.pred.5 <- knn(train = train.s, test.s.2, cl = train.s.y, k=5)
173.        getEvaluationMetrics(table(knn.pred.5, test.s.y))
174.
175.        knn.pred.10 <- knn(train = train.s, test.s.2, cl = train.s.y, k=10)
176.        getEvaluationMetrics(table(knn.pred.10, test.s.y))
177.
178.        numFolds <- trainControl(method = "cv", number = 10)
179.        kGrid <- expand.grid(.k = seq(3,25,2))
180.        set.seed(123)
181.
182.        knn.cv <-
     train(Churn ~ ., data = train.s.2, method = "knn",trControl = numFolds, tuneGrid = kGri
    d)
183.        knn.cv
184.
185.        # 17
186.        knn.pred.17 <- knn(train = train.s, test.s.2, cl = train.s.y, k=17)
187.        getEvaluationMetrics(table(knn.pred.17, test.s.y))
188.
189.        ######################################### SVM ################################
    ########
190.
191.        attach(churn)
192.        library(e1071)
193.
194.        test.svm <- test.s.2
195.        test.svm$Churn <- test.s.y
196.
197.        # Kreiranje modela
198.        svm.1 <- svm(Churn ~., data = train.s.2, type = "C-
    classification", kernel = "linear")
199.        svm.1
200.
201.        # Predikcije
202.        svm.1.pred <- predict(svm.1, test.svm)
203.        head(svm.1.pred)
204.
205.        svm.1
206.
207.        # Matrica konfuzije
208.        svm.1.cm <- table(predicted = svm.1.pred, true = test.svm$Churn)
209.        svm.1.cm
210.
211.        # Evaluacione metrike
212.        getEvaluationMetrics(svm.1.cm)
```