# Analiza karakteristika korisnika osiguravajućih usluga

# Apstrakt

Ovaj rad obuhvata analizu karakteristika korisnika osiguravajućih usliga njihovom klasifikacijom u odnosu na to da li su kupili ili ne osiguranje za karavan vozilo. Analiza osiguranika je neophodni segment rada bilo kog osiguravajućeg društva kako zbog identifikacije i smanjenja šansi za prevare kod različitih tipova osiguranja, tako i zbog redukcije troškova marketinga pomoću efikasne promocije. Projekat na osnovu koga je nastao ovaj rad nastao je u vidu saradnje kompanije *msgglobalsolutions* i katedre za upravljanje sistemima na Fakultetu organizacionih nauka. Korišćen je skup podataka Insurance Company Benchmark (COIL 2000) preuzet sa UCI machine learnig repository-ja (<http://kdd.ics.uci.edu/databases/tic/tic.html>). Podaci iz ovog skupa podataka su pretprocesirani čime se došlo do odgovarajućih ulaznih podataka za naš klasifikator, koji se pre svega tiču socio-demografskih karakteristika, kao i korišćenja drugih već postojećih osiguranja. Cilj ovog rada je da se segmentira deo već postojećih korisnika koji bi bio zainteresovan za kupovinu osiguranja za karavan vozila, putem različitih metoda klasifikacije.

# Uvod

Sektor osiguranja se sastoji od osiguravajućih društava čiji je osnovna uloga da svojim korisnicima pomognu da upravljaju rizikom poštujući formu polise osiguranja. Osnovni koncept osiguranja je da jedna strana, u ovom slučaju osiguravajuće društvo, garantuje isplatu ukoliko se ostvari određeni nepovoljni događaj, opisan u polisi osiguranja. U međuvremenu, druga strana, osiguranik, plaća male iznose tokom propisanog perioda vremena u zamenu za zaštitu od potencijalnih nepovoljnih događaja u budućnosti (<https://www.investopedia.com/ask/answers/051915/how-does-insurance-sector-work.asp#types-of-insurance-companies>).

Društva za osiguranje imaju i investicionu ulogu. Zakonom je takođe propisano na koji način društva za osiguranje deo prikupljenih premija mogu ulagati u hartije od vrednosti, nekretnine, depozite, itd. kako bi ostvarila veće prihode i obezbedila stabilno, pozitivno i dugoročno poslovanje (<http://www.skontajosiguranje.com/o-osiguranju.php>).

Svakako da je osnovni cilj svakog osiguravajućeg društva maksimiziranje profita, a samim tim i što veći broj korisnika osiguranja. Jedan od najefikasnijih načina za povećanje broja korisnika u bilo kojoj uslužnoj ili proizvodnoj kompaniji je marketing strategija koja se naziva unakrsna prodaja (eng. cross-selling). Unakrsna prodaja podrazumeva podsticanje postojećih kupaca da uz jedan proizvod ili uslugu kupe drugi srodan ili komplementaran proizvod ili uslugu s ciljem da se poveća prodaja i produbi odnos sa kupcem (<https://marketingfancier.com/unakrsna-prodaja-ili-cross-selling/>).

U 2000. godini evropska osiguravajuća kompanija koja nudi mnoštvo različitih vrsta osiguranja suočila se sa izazovom unakrsne prodaje kada se kompanijin najnoviji tip osiguranja za karavan vozilo pokazao razočaravajućim u pogledu prodaje. Marketing sektor kompanije je smatrao da bi podsticanje već postojećih osiguranika da kupe osiguranje za karavan značajno unapredilo prodaju. Međutim, glavno pitanje je bilo kako targetirati korisnike koji bi bili zainteresovani za ovaj tip osiguranja među nekoliko hiljada osiguranika i naj način ostvariti što veći prihod uz što manje troškove marketinga.

Jedan od načina za targetiranje osiguranika jeste korišćenje metoda mašinskog učenja. Mašinsko učenje predstavlja deo nauke koji omogućava računarima da uče, bez potrebe da za to budu eksplicitno programirani (Samuel, 1959, str. 210-229). Takođe, mašinko učenje može da se tumači kao kao sposobnost softverskog sistema da generalizuje na osnovu prethodnog iskustva i da potom koristi kreirane generalizacije kako bi pružio odgovore na pitanja koja se tiču entiteta/pojava koje pre nije sretao (Jovanović, 2016, str. 3).

Kao i u mnogim drugim proizvodnim ili uslužnim kompanijama, tako i u osiguravajućim društvima metode mašinskog učenja koje se tiču klasifikacije korisnika u one koji su zainteresovani za kupovinu određenog proizvoda odnosno korišćenje određene usluge i one koji nisu se jako često koriste. Pravilno klasifikovanje potencijalnih osiguranika može biti izrazito korisno u fazi promocije i na taj način u mnogome umanjiti troškove marketinga.

Cilj ovog rada je da se izvrši binarna klasifikacija postojećih osiguranika, odnosno podela u dve grupe na osnovu karakteristika primećenih u podacima a koja će dovesti do zaključka da li je određeni korisnik zainteresovan za kupovinu karavan osiguranja.

# Opis konkretnog problema

Kao što je već pomenuto, projekat na osnovu koga je nastao ovaj rad bazira se na problemu binarne klasifikacije osiguranika jedne evropske osiguravajuće kuće, na osnovu toga da li žele da kupe osiguranje za karavan vozilo ili ne.

Skup podataka korišćen u samom eksperimentu obezbedila je holandska kompanija koja se bavi data mining-om, *Sentient Machine Research.* Podaci su podeljeni su na trening skup (5822 opservacije) koji sadrži podatke o izlaznoj promenljivoj i test skup (4000 opservacija) koji ne sadrži podatke o izlaznoj promenljivoj.

Svaka opservacija, odnosno red u skupu podataka sačinjena je od 86 atributa (kolona), od koji se prvih 43 odnose na socio-demografske karakterstike pojedinačnih osiguranika, a narednih 43 na korišćenje već postojećih vrsta osiguranja kompanije. Socio-demografske karakteristike su izvedene iz poštanskog broja (eng. zip code) kome pripada dati korisnik, tako da su vrednosti ovih promenljivih identične za one korisnike koji dolaze iz područja sa istim poštanskim brojem. Poslednja kolona predstavlja izlaznu promenljivu čije su vrednosti 0 ili 1, gde je 0 indikator da korisnik ne želi da kupi uslugu karavan osiguranja, a 1 indikator da želi. Kako je cilj kompanije da se što tačnije predvide svi potencijalni kupci karavan osiguranja kako bi se promocija usmerila isključivo na njih i na taj način smanjili troškovi, za pozitivnu klasu izlazne promenljive koristiće se kupci karavan osiguranja, dok će se za negativnu koristiti oni korisnici koji nisu zainteresovani za ovaj tip osiguranja.

Ceo projekat je odrađen u programskom jeziku R, kao jednom od najadekvatnijih programskih jezika za potrebe mašinskog učenja. Programsko okruženje korišćeno za rad je RStudio.

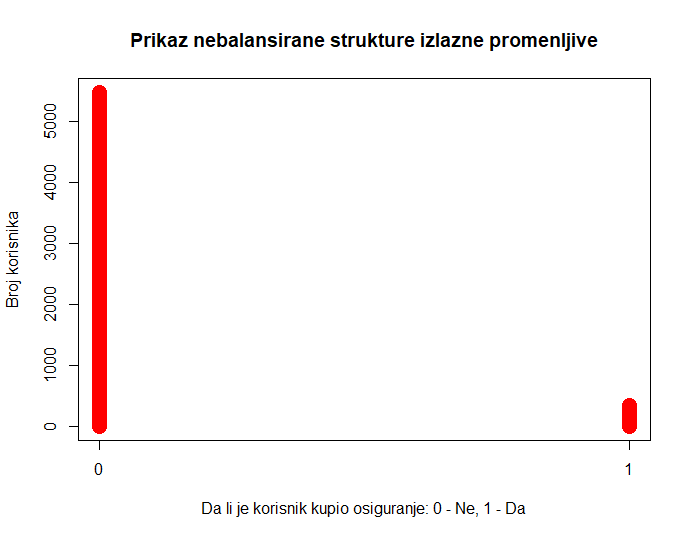
# Eksperiment

U narednom delu nalazi se kompletan prikaz eksperimenta, odnosno sve faze samog projekta. Te faze obuhvataju prikupljanje podataka, njihovu analizu i vizuelizaciju, određivanje skupa atributa koji će biti korišćeni i na kraju sam rad na klasifikatoru.

## Struktura podataka

S obzirom da je za potrebe ovog projekta korišćen gotov skup podataka, prvi zadatak je bio učitati podatke u samo programsko okruženje. S obyirom da skup podataka sadrži veliki broj atributa, čak 86, prikazivanje rečnika podataka u ovom radu ne bi bilo pregledno, međutim on je dostupan na sledećem linku  <http://kdd.ics.uci.edu/databases/tic/dictionary.txt>. Prva karakteristika podataka koja se može zapaziti je da su svih 86 promenljivih celobrojni tipovi (integer), dok je u rečniku podataka objašnjeno koja skala je korišćena za svaku pojedinačnu promenljivu.

Neizbalansiranost skupa podataka predstavlja čestu situaciju prilikom rešavanja problema mašinskog učenja. Neizbalansiranost skupa podataka predstavlja nejednaku distribuciju klasa pojedinih promenljivih (<https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb>). Ovo je čest slučaj u rešavanju problema mašinskog učenja, pogotovu kada se radi o problemima klasifikacije (npr. u medicini kada se ispituje da li osoba ima neku retku bolest gde će po pravilu samo mali procenat osoba stvarno biti zaražen). U ovom projektu ispituje se da li će korisnik nekog od osiguranja biti zainteresovan i za kupovinu karavan osiguranja. Logičkim razmišljanjem može se zaključiti da će mali broj korisnika želeti da kupi ovo osiguranje, pre svega zbog činjenice da mali broj korisnika zapravo i poseduje karavan vozilo. Da je ovo razmišljanje ispravno možemo videti na jednostavnom grafiku koji pokazuje da samo 6.4% korisnika u trening setu poseduje osiguranje za karavan vozilo.



Ukoliko se rešava problem klasifikacije, a izlazna promenljiva je nebalansirana, vrlo lako je moguće napisati “lažne” algoritme koji će svakom novom slučaju dodeliti najčešće pojavljivanje vrednosti izlazne klase i na taj način postići veliki procenat tačnosti. Naravno, očekuje se i od svih drugih algoritma da ostvaruju veliki procenat tačnosti (eng. accuracy) takođe zbog ovog disbalansa, jer će tačno klasifikovati veliki procenat observacija negativne klase. Međutim, to ne treba da nas zavara, jer će se uspešnost klasifikatora ogledati pre svega u drugim metrikama koje računaju udeo pozitivne klase što su “sensitivity” (odnos tačno predviđenih pozitivnih varijabli i svih varijabli koje su stvarno pozitivne) i “PPV” (odnos tačno predviđenih pozitivnih varijabli i svih predviđenih pozitivnih varijabli). Ove metrike dace nam odličan uvid u to Koliko dobro predviđamo pozitivne opservacije i biće jako korisne u evaluaciji našeg modela.

Međutim, postoje i određene tehnike koje rešavaju probem nebalansiranog skupa podataka. Najpoznatije među njima su “sampling” tehnike i to “undersampling” i “oversampling”. Prva predstavlja proces nasumičnog brisanja opservacija iz dominantne klase (u našem slučaju negativne klase koja se odnosi na korisnike koji nisu zainteresovani za kupovinu karavan osiguranja) radi izjednačavanja broja opservacija obe klase. “Oversampling” predstavlja process nasumičnog generisanja uzorka iz opservacija nedominantne klase takođe radi izjednačavanja broja opservacija obe klase. Postoji više različitih “oversampling” tehnika. Najpoznatije su “random oversampling” i “SMOTE” (synthetic minority oversampling technique). Prva se odnosi na nasumično biranje opservacija nedominantne klase i njihovo ponavljanje do izjednačavanja broja opservacija obe klase. “SMOTE” bira k najbližih suseda za svaku opservaciju nedominantne klase, spaja ih sa datom opservacijom I na tim putanjama generiše nove slučajeve (<https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb>).

U ovom radu, svaki model ćemo testirati na tri skupa podataka i to nebalansiranom skupu, “undersampling” skupu I “oversampling” skupu I pokušati da utvrdimo da li I na koji način “sampling” tehnike poboljšavaju korišćene modele. Međutim, pre kreiranja novih skupova podataka uvek je poželjno “osloboditi” se viška atributa, odnosno onih atributa za koje se smatra da ne doprinose samoj klasifikaciji nekom od metoda redukcije atributa. Pre svega toga, nije lose malo bolje se upoznati sa skupom podataka.

## Eksploratorna analiza podataka

Nakon prikupljanja podataka, uobačajeni korak predstavlja upoznavanje sa podacima i sagledavanje šta predstavljaju same promenljive i na koji način mogu uticati na rešavanje postavljenog problema. Poželjno je upoznati se njihovim glavnim karakteristikama, što je najčešće slučaj uz pomoć vizuelnih metoda. U ovom delu projekta nije neophodno koristiti statističke modele, već je prioritet razumeti podatke izvan formalnih modela i testova nad postavljenim hipotezama. Stoga, sledeći korak u radu predstavlja eksploratorna analiza podataka. Treba napomenuti da skup podataka korišćen u ovom projektu ne sadrži nedostajuće vrednosti, tako da podatke možemo u startu smatrati za adekvatno pripremljenje za analizu.

Eksploratorna analiza podataka predstavlja proces inicijalnog istraživanja podataka radi otkrivanja paterna, uočavanja anomalija, testiranja hipoteza i provere pretpostavki uz pomoć summary statistike i grafičke reprezentacije (<https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>).

### Univarijaciona eksploratorna analiza

Univariaciona analiza je najjednostavniji oblik eksploratorne analize podataka. “Uni“ znači “jedan” što zapravo znači da se ovaj tip analize bavi isključivo pojedinačnim promenljivama, a ne uzrocima i vezama (za razliku od regresije) i njena glavna svrha je da opiše podatke, odnosno da pronađe paterne u njima (<https://www.statisticshowto.com/univariate/>).

Kako 81 od 86 promenljivih iz skupa podataka spada u jednu od tri kategorije po kojima su predstavljene promenljive, u ovom delu ćemo se pozabaviti deskriptivnim statistikama svake od ove tri kategorije.

Prva kategorija obuhvata sociodemografske podatke čiji opseg varira od 0 do 9 gde svaki broj predstavlja određeni procentualni opseg zastupljenosti pojedinačne karakteristike u oblasti sa datim poštanskim brojem.

Tabela 1: Značenje vrednosti sociodemografskih atributa

|  |  |
| --- | --- |
| Broj | Procentualni opseg |
| 0 | 0% |
| 1 | 1 – 10% |
| 2 | 11 – 23% |
| 3 | 24 – 36% |
| 4 | 37 – 49% |
| 5 | 50 – 62% |
| 6 | 63 – 75% |
| 7 | 76 – 88% |
| 8 | 89 – 99% |
| 9 | 100% |

### TODO: Dodati tabelu 2 ###

Na osnovu tabele 2 lako se primeti da velika većina sociodemografskih atributa sadrži maksimalne opsege, odnosno sadrži opservacije i sa 0 i 100 procenata, što dosta govori o korisnika opservacija iz različitih sredina. Vidimo da samo tri promenljive odstupaju od ove osobine.

Druga kategorija obuhvata novčano ulaganje svakog pojedinačnog korisnika u svaku pojedinačnu već postojeću vrstu osiguranja. Opseg vrednosti ove grupe atributa takođe varira od 0 do 9 gde svaka vrednost predstavlja određeni opseg novčaniog ulaganja u pojedinačno osiguranje.

Tabela 3: Značenje vrednosti atributa za novčano ulaganje u osiguranja

|  |  |
| --- | --- |
| Broj | Novčani opseg |
| 0 | 0 f |
| 1 | 1 – 49 f |
| 2 | 50 – 99 f |
| 3 | 100 – 199 f |
| 4 | 200 – 499 f |
| 5 | 500 – 999 f |
| 6 | 1000 – 4999 f |
| 7 | 5000 – 9999 f |
| 8 | 10000 – 19999 f |
| 9 | Preko 20000 f |

### TODO: Dodati tabelu 4 ###

Iz tabele 4 može se zaključiti da za razliku od prve kategorije, mali broj atributa obuhvata maksimalan opseg 0-9. Razlog tome je što samo tri osiguranja koji ima bar jednog korisnika koji je uložio preko 20000 novčanih jedinica u to osiguranje i to su kao što se i moglo pretpostaviti životno osiguranje, osiguranje na automobil i kamion. Takođe možemo primetiti da su standardne devijacije ovih atributa dosta manje od sociodemografskih atributa, što i ne treba da čudi jer veliki broj korisnika nije kupovao određena pojedinačna osiguranja a samim tim nije ni ulagao novčana sredstva u njih. Treba napomenuti da od ovog zaključka odtupa osiguranje za automobile sa standardnom devijacijom bliz 3, jer ipak veliki broj korisnika kupuje ovaj tip osiguranja, a s obzirom na različite finansijske mogućnosti opseg ulaganja je takođe veliki. Međutim, ono što je možda i najbitnije za zaključiti iz deskriptivne statistike ovih promenljivih je da samo osiguranje za automobile i od požara imaju medijanu različitu od nule odakle proističe da ogroman broj korisnika uopšte nije ni kupovao druga osiguranja.

Poslednja kategorija promenljivih odnosi se na broj polisa svakog pojedinačnog osiguranja. Opseg vrednosti ove kategorije atributa varira od 0 do 12 gde svaki broj predstavlja ujedno i broj polisa konkretnog osiguranja.

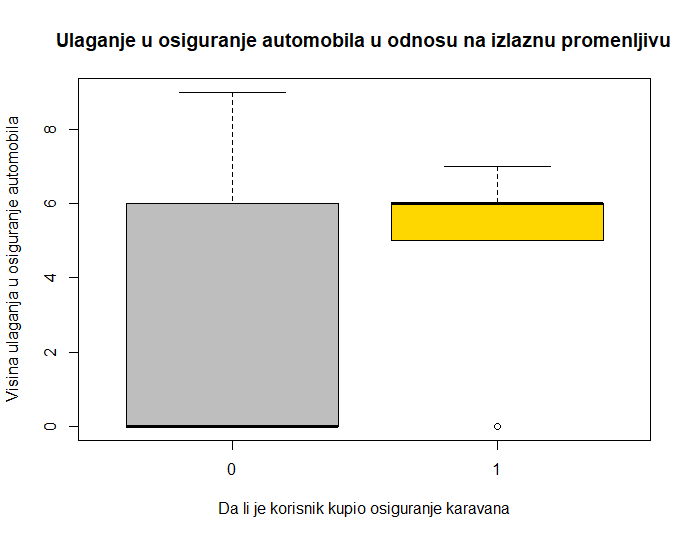
### TODO: Dodati tabelu 5 ###

Na osnovu tabele 5 vidimo da su standardne devijacije ponovo jako niske, ovoga puta sve ispod 1, dok se na prva dva mesta takođe nalaze osiguranje za automobil i od požara. Samo osiguranje na automobile ima korisnika sa 12 polisa, dok čak 4 tipa osiguranja nemaju ni je dnog korisnika sa brojem polisa većeg od 1. Takođe, kao i kod promenljivih koje se odnose na ulaganja u osiguranja, samo 2 tipa osiguranja imaju medijanu različitu od nule i to ponovo osiguranje za automobile i od požara.

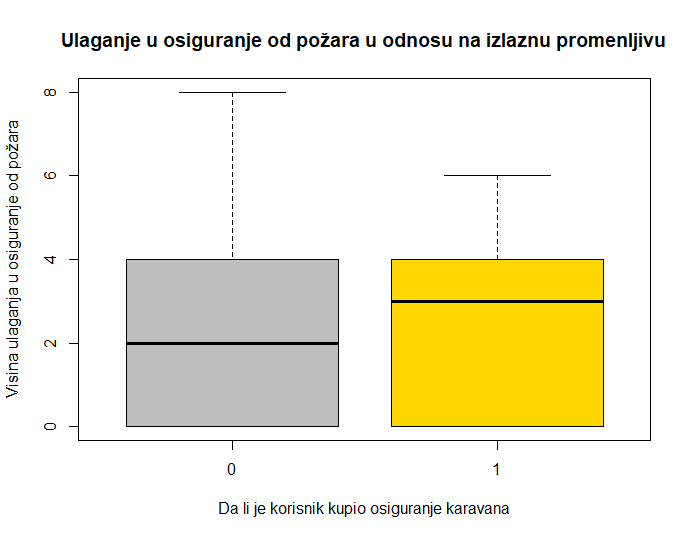
Dakle, lako se uočava da se sociodemografski atributi u mnogome razlikuju od onih koji se odnose na osiguranja. Paterni koji se uočavaju u drugoj i trećoj kategoriji promenljivih dovode do pitanja da li ti atributi uopšte imaju značaja u određivanju vrednosti izlazne promenljive, pre svega činjenica da u obe pomenute kategorije samo dva tipa osiguranja imaju medijanu različitu od nule, a samo tri tipa treći kvartil različit od nule. Međutim, da bi se sa većom dozom sigurnosti moglo reći da većina ovih promenljivih nije značajna za problem klasifikaicje korisno je poslužiti se metodama redukcije atributa što će i biti tema u kasnijem delu rada.

### Multivarijaciona eksploratorna analiza

Za razliku od univarijacione analize koja se bavi pojedinačnim promenljivama, multivarijaciona eksploratorna analiza se bavi vezama između promenljivih. Analizirajući pojedinačne promenljive pokazalo se da najveći broj atributa koji se odnose na pojedinačna osiguranja zapravo sadrži ogromne procente polja sa vrednošču nula. Odstupanje od ovog paterna moglo se primetiti kod osiguranja za automobil i osiguranja od požara, tako da će ovaj deo rada biti posvećen pre svega vizuelnom prikazu odnosa ova dva atributa sa izlaznom promenljivom.



Sa slike 1 može se primetiti velika razlika ponašanja korisnika u pogledu ulaganja u osiguranje automobile i kupovine osiguranja karavana. Postoji samo jedan izuzetak u vidu korisnika koji nije osigurao automobile a jeste caravan vozilo. Zapaža se da velika razlika u medijani u ove dve grupe. Više od pola korisnika koji nisu kupili caravan osiguranje, nisu kupili ni osiguranje automobila, odnosno nisu ulagali u njega, dok polovina korisnika koji su osigurali caravan, uložili su preko hiljadu novčanih jedinica u osiguranje automobila. Ova raznolikost govori da će promenljiva koja se odnosi na ulaganje u osiguranje automobile definitivno imati značaja prilikom klasifikacije.



Pogledajmo sada kakav je odnos ulaganja u osiguranje od požara I izlazne promenljive. Može se primetiti da razlike između dve grupe nisu toliko velike kao u prethodnom slučaju. Postoje korisnici koji nisu kupovali ni jednu od ove dve vrste osiguranja, međutim interesantno je videte Koliko novca ulažu u osiguranje od požara obe grupe osiguranika. Medijana grupe koja je kupila caravan osiguranje je nešto veća, što nas dovodi do zaključka da oni koji su kupili osiguranje karavana pretenduju da daju nešto veće sume novca za osiguranje od požara, ali interesantno je da su to sume do 5000 novčanih jedinica, dok oni koji se odluče da daju mnogo veće sume novca u osiguranje od požara, u iznosu od 5 do 20 hiljada, ne odlučuju se za kupovinu osiguranja karavana. Sve ovo govori u prilog tome da za razliku od prethodno analiziranog atributa, sada nije moguće sa sigurnošću reći Koliko značaja ima ova promenljiva u određivanju izlazne.

Da bismo malo detaljnije proučili kako se kreću brojke kada poredimo zavisnost dve promenljive možemo se poslužiti barchart-ovima.

# Učitavanje data set-a za treniranje

train.data **<-** read.csv**(**"ticdata2000.csv.xls",header **=** F, stringsAsFactors **=** **FALSE)**

train.data**[**1,1**]** **<-** 33

colnames**(**train.data**)** **<-** c**(**"Customer.Subtype", "Number.of.houses", "Avg.size.household","Avg.age","Customer.main.type","Roman.catholic","Protestant","Other.religion", "No.religion", "Married", "Living.together", "Other.relation","Singles", "Household.without.children","Household.with.children", "High.level.education", "Medium.level.education", "Lower.level.education", "High.status", "Entrepreneur","Farmer","Middle.management","Skilled.labourers", "Unskilled.labourers", "Social.class.A","Social.class.B1","Social.class.B2", "Social.class.C", "Social.class.D", "Rented.house", "Home.owners", "1.car", "2.cars", "No.car","National.Health.Service","Private.health.insurance","Income.less.then.30K","Income.in.30.45K", "Income.in.45.75K", "Income.in.75.122K", "Income.more.then.123K", "Average.income", "Purchasing.power.class", "Cont.private.third.party.insr", "Cont.third.party.insr.firms", "Cont.third.party.insr.agriculture", "Cont.car.pol", "Cont.delivery.van.pol", "Cont.motorcycle.scooter.pol", "Cont.lorry.pol", "Cont.trailer.pol", "Cont.tractor.pol", "Cont.agricultural.machines.pol", "Cont.moped.pol", "Cont.life.insr", "Cont.private.accident.insr.pol", "Cont.family.accidents.insr.pol", "Cont.disability.insr.pol", "Cont.fire.pol", "Cont.surfboard.pol","Cont.boat.pol",

"Cont.bicycle.pol",

"Cont.property.insr.pol",

"Cont.social.security.insr.pol",

"Num.of.private.third.party.insur",

"Num.of.third.party.insr.firms",

"Num.of.third.party.insr.agriculture",

"Num.of.car.pol",

"Num.of.delivery.van.pol",

"Num.of.motorcycle.scooter.pol",

"Num.of.lorry.pol",

"Numb.of.trailer.pol",

"Num.of.tractor.pol",

"Num.of.agricultural.machines.pol",

"Num.of.moped.pol",

"Number.of.life.insr",

"Number.of.private.accident.insr.pol",

"Number.of.family.accidents.insr.pol",

"Number.of.disability.insr.pol",

"Number.of.fire.pol","Num.of.surfboard.pol", "Number.of.boat.pol","Num.of.bicycle.pol", "Num.of.property.insr.pol", "Num.of.social.security.insr.pol", "Num.of.mobile.home.policies"**)**

# Pregled data set-a za treniranje

View**(**train.data**)**

# Učitavanje data set-a za testiranje

test.data **<-** read.csv**(**"ticeval2000.csv.xls",header **=** F, stringsAsFactors **=** **FALSE)**

test.data**[**1,1**]** **<-** 33

colnames**(**test.data**)** **<-** colnames**(**train.data**[-**86**])**

# Učitavanje izlazne varijable za data set- za treniranje, koju ćemo koristiti pri proveri našeg modela

test.target **<-** read.csv**(**"tictarget.csv",header **=** F, stringsAsFactors **=** **FALSE)**

colnames**(**test.target**)** **<-** colnames**(**train.data**[**86**])**

test.data **<-** cbind**(**test.data,test.target**)**

Smatrali smo da ukoliko naš model obuhvati svih 85 varijabli da može doći do toga da naš model ne predviđa dobro, odnosno da mu bude slaba moć generalizacij (overfitting problem). Iz ovog razloga smo odlučili da redukujemo broj varijabli koje će učestvovati u predikciji. Daljom analizom i korišćenjem korelacionih matrica, došli smo do zaključka koje kolone bismo to mogli da izbacimo.

summary**(**train.data**[**,65**:**85**])**

#KORELACIJA

install.packages**(**"corrplot"**)**

library**(**corrplot**)**

corr\_matrix**<-** cor**(**train.data**[**,c**(**16**:**29,37**:**42**)])**

corrplot**(**corr\_matrix, type **=** "upper", diag **=** F,tl.pos **=** **NULL**, tl.cex **=** 0.3, tl.col **=** "red", tl.offset **=** 0.4, tl.srt **=** 90**)**

data **<-** rbind**(**train.data,test.data**)**

corr\_matrix\_2**<-** cor**(**data**[**,c**(**16**:**29,37**:**42**)])**

corrplot**(**corr\_matrix\_2, type **=** "upper", diag **=** F,tl.pos **=** **NULL**, tl.cex **=** 0.3, tl.col **=** "red", tl.offset **=** 0.4, tl.srt **=** 90**)**

zeros\_train **<-** colSums**(**train.data**[**,c**(**44**:**85**)]==**0**)/**nrow**(**train.data**[**,c**(**44**:**85**)]==**0**)\***100

zeros\_sum **<-** colSums**(**data**[**,c**(**44**:**85**)]==**0**)/**nrow**(**data**[**,c**(**44**:**85**)]==**0**)\***100

zeros\_train

zeros\_sum

which**(**zeros\_train**<**90**)**

which**(**zeros\_sum**<**90**)**

corr\_matrix\_2**<-** cor**(**train.data**[**,c**(**44**:**64,86**)])**

corrplot**(**corr\_matrix\_2, type **=** "upper", diag **=** F,tl.pos **=** **NULL**, tl.cex **=** 0.3, tl.col **=** "red", tl.offset **=** 0.4, tl.srt **=** 90**)**

Ispitali smo korelaciju izmedju pojedinih sociodemografskih karakteristika za koje smo smatrali da bi mogle da budu visoko korelisane, što se i ispostavilo kao slučaj kod varijabli koje se tiču obrazovanja, visine prihoda i pripadnosti određenoj socijalnoj klasi. Na osnovu ovog smo odlučili da izbacimo varijable koje se odnose na pripadnost socijalnoj klasi jer su one bile u najmanjoj korelaciji sa izlaznom varijablom u odnosu na ostale pomenute.

Varijable koje se odnose na posedovanje osiguranja podeljene su u dve grupe, gde se prva grupa odnosi na količinu novca uloženu na svaku od specifičnih vrsta osiguranja, dok se u drugoj grupi nalazi broj kupljenih osiguranje od svake ove vrste. Što se tiče ovih grupa varijabli koje se odnose na posedovanja osiguranja, primetili smo prethodnom analizom da ima dosta kolona čije opservacije većim delom imaju 0. Iz tog razloga smo proverili procentualni udeo nula u svakoj od ovih kolona i utvrdili da je u malom broju njiih taj procentualni udeo bio manji od 95% zbog čega smo zaključili da ćemo za obe grupe zadržati samo te kolone, a ostale agregirati u jednu varijablu.

#Dobijanje agregirane varijable za uloženi novac u ostala osiguranja

train.data.other.contribution **<-** train.data**[**, 44**:**64**]**

train.data.other.contribution**$**Cont.private.third.party.insr **<-** **NULL**

train.data.other.contribution**$**Cont.car.pol **<-** **NULL**

train.data.other.contribution**$**Cont.fire.pol **<-** **NULL**

train.data.other.contribution**$**Cont.boat.pol **<-** **NULL**

train.data.other.contribution **<-** as.data.frame**(**apply**(**train.data.other.contribution, 1, sum**))**

#Dobijanje agregirane varijable za broj kupljenih ostalih osiguranja

train.data.other.number.of.pol **<-** train.data**[**, 65**:**85**]**

train.data.other.number.of.pol**$**Num.of.private.third.party.insur **<-** **NULL**

train.data.other.number.of.pol**$**Num.of.car.pol **<-** **NULL**

train.data.other.number.of.pol**$**Number.of.fire.pol **<-** **NULL**

train.data.other.number.of.pol**$**Number.of.boat.pol **<-** **NULL**

train.data.other.number.of.pol **<-** as.data.frame**(**apply**(**train.data.other.number.of.pol, 1, sum**))**

train.data.new **<-** train.data**[**,**-**c**(**6**:**9, 42, 16**:**18, 25**:**29, 45, 46, 48**:**58, 60, 62**:**64, 66, 67, 69**:**79, 81, 83**:**85**)]**

train.data.new**$**Other.contribution **<-** train.data.other.contribution

train.data.new**$**Number.of.other.policies **<-** train.data.other.number.of.pol

#Za data set za testiranje dobijanje agregirane varijable za uloženi novac u ostala osiguranja

test.data.other.contribution **<-** test.data**[**, 44**:**64**]**

test.data.other.contribution**$**Cont.private.third.party.insr **<-** **NULL**

test.data.other.contribution**$**Cont.car.pol **<-** **NULL**

test.data.other.contribution**$**Cont.fire.pol **<-** **NULL**

test.data.other.contribution**$**Cont.boat.pol **<-** **NULL**

test.data.other.contribution **<-** as.data.frame**(**apply**(**test.data.other.contribution, 1, sum**))**

#Za data set za testiranje dobijanje agregirane varijable za broj kupljenih ostalih osiguranja

test.data.other.number.of.pol **<-** test.data**[**, 65**:**85**]**

test.data.other.number.of.pol**$**Num.of.private.third.party.insur **<-** **NULL**

test.data.other.number.of.pol**$**Num.of.car.pol **<-** **NULL**

test.data.other.number.of.pol**$**Number.of.fire.pol **<-** **NULL**

test.data.other.number.of.pol**$**Number.of.boat.pol **<-** **NULL**

test.data.other.number.of.pol **<-** as.data.frame**(**apply**(**test.data.other.number.of.pol, 1, sum**))**

test.data.new **<-** test.data**[**,**-**c**(**6**:**9, 42, 16**:**18, 25**:**29, 45, 46, 48**:**58, 60, 62**:**64, 66, 67, 69**:**79, 81, 83**:**85**)]**

test.data.new**$**Other.contribution **<-** test.data.other.contribution

test.data.new**$**Number.of.other.policies **<-** test.data.other.number.of.pol

Nakon ovih analiza redukovani data set sadrži 41. varijablu.