Za potrebe primene različitih tehnika mašinskog učenja koristićemo sledeći data set: Insurance Company Benchmark (COIL 2000) Data Set.Ovaj data set sadrži podatke vezane za korisnike različitih osiguranja, tačnije njihove sociodemografske karakteristike kao i podatke vezane za direktno korisćenje nekih od usluga osiguranja. Ovi podaci su tako prikupljeni da nam pomognu da na osnovu njih predvidimo da li će korisnik biti zainteresovan za kupovinu osiguranja za karavan vozilo.

Svaka opservacija, odnosno red u data setu sačinjen je od 86 varijabli (kolona), od kojih se prvih 43 (1-43) odnose na sociodemografske podatke, a narednih 43 (44-86) na njihove usluge različitih vrsta osiguranja. Sociodemografske karakteristike su izvedene iz zip koda kome pripada dati korisnik, dok se naredne varijable odnose na njegove usluge osiguranja koje već koristi. Poslednja kolona je target kolona, odnosno izlazna varijabla čije su vrednosti 1 ili 0, gde je 1 indikator da će korisnik biti potencijalno zainteresovan i za uslugu karavan osiguranja, a 0 indikator da neće.

Podaci su preuzeti iz holandske kompanije za data mining Sentient Machine Research i zasnovani su na realnom poslovnom problemu u svetu. Podaci su podeljeni na skup podatka za test i skup podataka za trening. Podaci za trening sadrže 5822 opservacije, u okviru kojih sadrže i informaciju da li korisnik trenutno poseduje osiguranje za karavan, odnosno podatke o target varijabli. Podaci za test sadrže 4000 opservacija. Naš zadatak u test setu je da odredimo da li će korisnik uzeti i uslugu karavan osiguranja sa najvećom mogućom preciznošću. Pozitivna klasa će nam biti oni koji će predikcijom upasti u grupu onih koji ne žele da kupe ovo osiguranje da bismo njih pokušali da zainteresujemo dodatnim marketingom.

Odlučili smo da za rešavanje datog problema koristimo RStudio i počeli smo od učitavanja skupa podataka iz csv fajlova i njihovo čuvanje u okviru novih promenljivih. Na ovaj način smo učitali i trening i test set i imenovali njihove kolone.

# 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.