Sadržaj

[Analiza karakteristika korisnika osiguravajućih usluga 2](#_Toc39917022)

[Apstrakt 2](#_Toc39917023)

[Uvod 3](#_Toc39917024)

[Opis konkretnog problema 4](#_Toc39917025)

[Analiza i pretprocesiranje podataka 4](#_Toc39917026)

[Struktura podataka 5](#_Toc39917027)

[Eksploratorna analiza podataka 6](#_Toc39917028)

[Univarijaciona eksploratorna analiza 7](#_Toc39917029)

[Multivarijaciona eksploratorna analiza 10](#_Toc39917030)

[Redukcija atributa 12](#_Toc39917031)

[Filter metode 13](#_Toc39917032)

[Obmotavajuće metode – Boruta algoritam 15](#_Toc39917033)

# Analiza karakteristika korisnika osiguravajućih usluga

# Apstrakt

Ovaj rad se bavi analizom karakteristika korisnika osiguravajućih usluga njihovom klasifikacijom u odnosu na to da li su kupili osiguranje karavan vozila. 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 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 karavan vozila, putem različitih metoda klasifikacije.

# Uvod

Sektor osiguranja se sastoji od osiguravajućih društava čija 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](https://www.investopedia.com/ask/answers/051915/how-does-insurance-sector-work.asp" \l "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 karavan vozila pokazao razočaravajućim u pogledu prodaje. Marketing sektor kompanije je smatrao da bi podsticanje već postojećih osiguranika da kupe osiguranje karavan vozila značajno unapredilo prodaju. Međutim, glavno pitanje je bilo kako segmentirati 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 računarske inteligencije. Računarska inteligencija predstavlja deo nauke koji omogućava računarima da uče, bez potrebe da za to budu eksplicitno programirani (Samuel, 1959, str. 210-229). Takođe, računarska inteligencija 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 računarske inteligencije 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 osiguranja karavan vozila.

# 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.

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 (verzija 3.6.2), kao jednom od najadekvatnijih programskih jezika za potrebe računarske inteligencije. Programsko okruženje korišćeno za rad je RStudio (verzija 1.2.5033).

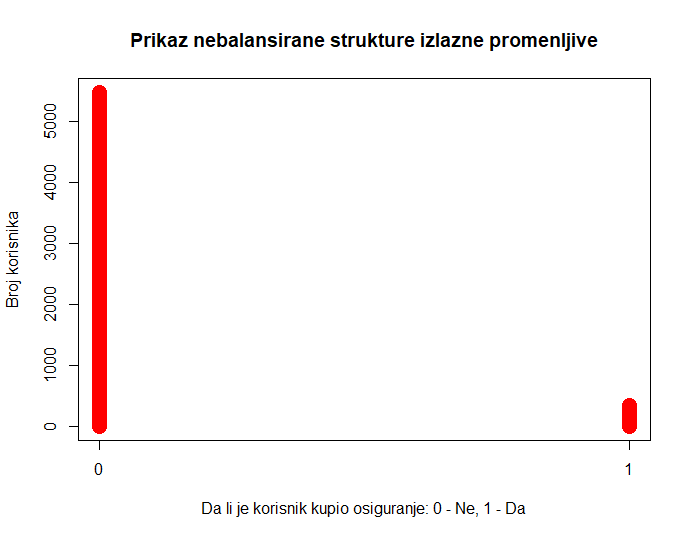
# Analiza i pretprocesiranje podataka

U narednom delu akcenat će biti na analizi i pretprocesiranju podataka koji su neophodni da bi se originalni skup podataka bolje razumeo, a potom i obradio i da bi se na kraju od njega dobio redukovani na kom će se vršiti sam eksperiment.

## 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 obzirom 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 može se videti na jednostavnom grafiku koji pokazuje da samo 6.4% korisnika u trening setu poseduje osiguranje za karavan vozilo.



Slika 1: Odnos korisnika koji su kupili osiguranje karavana i onih koji nisu

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 (eng. accuracy). Naravno, očekuje se i od svih drugih algoritma da ostvaruju veliki procenat tačnosti takođe zbog ovog disbalansa, jer će tačno klasifikovati veliki procenat opservacija negativne klase. Međutim, to ne treba da zavara, jer će se uspešnost klasifikatora ogledati pre svega u drugim metrikama koje računaju udeo pozitivne klase što su senzitivnost (eng. Sensitivity - odnos tačno predviđenih pozitivnih promenljivih i svih varijabli koje su stvarno pozitivne) i preciznost (eng. Percision - odnos tačno predviđenih pozitivnih promenljivih i svih predviđenih pozitivnih promenljivih). Ove metrike daće odličan uvid u to koliko dobro se predviđaju pozitivne opservacije i biće jako korisne u evaluaciji 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 loše 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 deskriptivne statistike i grafičke reprezentacije (<https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>).

### Univarijaciona eksploratorna analiza

Univarijaciona analiza je najjednostavniji oblik eksploratorne analize podataka. “Uni“ znači “jedan” što zapravo govori 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% |

Tabela 2: Deskriptivne statistike socio-demografskih atributa

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | std | min | 25% | 50% | 75% | max |
| Roman.catholic | 0.701 | 1.015 | 0 | 0 | 0 | 1 | 7 |
| Protestant | 4.638 | 1.721 | 0 | 4 | 5 | 6 | 9 |
| Other.religion | 1.050 | 1.011 | 0 | 0 | 1 | 2 | 5 |
| No.religion | 3.263 | 1.606 | 0 | 2 | 3 | 4 | 9 |
| Married | 6.189 | 1.896 | 0 | 5 | 6 | 7 | 9 |
| Living.together | 0.873 | 0.962 | 0 | 0 | 1 | 1 | 7 |
| Other.relation | 2.287 | 1.711 | 0 | 1 | 2 | 3 | 9 |
| Singles | 1.887 | 1.779 | 0 | 0 | 2 | 3 | 9 |
| Household.without.children | 3.237 | 1.609 | 0 | 2 | 3 | 4 | 9 |
| Household.with.children | 4.303 | 1.984 | 0 | 3 | 4 | 6 | 9 |
| High.level.education | 1.485 | 1.646 | 0 | 0 | 1 | 2 | 9 |
| Medium.level.education | 3.307 | 1.723 | 0 | 2 | 3 | 4 | 9 |
| Lower.level.education | 4.592 | 2.280 | 0 | 3 | 5 | 6 | 9 |
| High.status | 1.899 | 1.814 | 0 | 0 | 2 | 3 | 9 |
| Entrepreneur | 0.403 | 0.787 | 0 | 0 | 0 | 1 | 5 |

U tabeli 2 prikazano je prvih 15 socio-demografskih atributa. Može se primeti da većina socio-demografskih atributa sadrži maksimalne opsege, odnosno sadrži opservacije i sa 0 i 100 procenata, što dosta govori o raznolikosti korisnika iz različitih sredina. 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 |

Tabela 4: Deskriptivne statistike atributa vezanih za ulaganje u osiguranja

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | std | min | 25% | 50% | 75% | max |
| Cont.private.third.party.insr | 0.765 | 0.957 | 0 | 0 | 0 | 2 | 3 |
| Cont.third.party.insr.firms | 0.039 | 0.357 | 0 | 0 | 0 | 0 | 6 |
| Cont.third.party.insr.agriculture | 0.074 | 0.508 | 0 | 0 | 0 | 0 | 4 |
| Cont.car.pol | 2.956 | 2.921 | 0 | 0 | 5 | 6 | 9 |
| Cont.delivery.van.pol | 0.055 | 0.566 | 0 | 0 | 0 | 0 | 7 |
| Cont.motorcycle.scooter.pol | 0.170 | 0.889 | 0 | 0 | 0 | 0 | 7 |
| Cont.lorry.pol | 0.009 | 0.238 | 0 | 0 | 0 | 0 | 9 |
| Cont.trailer.pol | 0.019 | 0.201 | 0 | 0 | 0 | 0 | 5 |
| Cont.tractor.pol | 0.094 | 0.604 | 0 | 0 | 0 | 0 | 7 |
| Cont.agricultural.machines.pol | 0.012 | 0.215 | 0 | 0 | 0 | 0 | 6 |

Tabela 4 prikazuje prvih 10 atributa koji govore o ulaganju novca u različita osiguranja. Može 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 socio-demografskih 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 odstupa osiguranje za automobile sa standardnom devijacijom blizu 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.

Tabela 5: Deskriptivne statistike atributa vezanih za broj različitih osiguranja

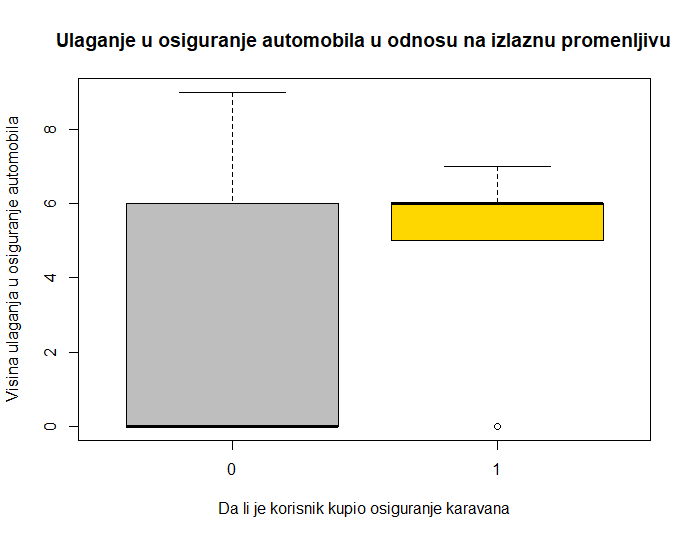
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | std | min | 25% | 50% | 75% | max |
| Num.of.private.third.party.insr | 0.400 | 0.492 | 0 | 0 | 0 | 1 | 2 |
| Num.of.third.party.insr.firms | 0.014 | 0.126 | 0 | 0 | 0 | 0 | 5 |
| Num.of.third.party.insr.agriculture | 0.021 | 0.144 | 0 | 0 | 0 | 0 | 1 |
| Num.of.car.pol | 0.557 | 0.609 | 0 | 0 | 1 | 1 | 12 |
| Num.of.delivery.van.pol | 0.011 | 0.130 | 0 | 0 | 0 | 0 | 5 |
| Num.of.motorcycle.scooter.pol | 0.040 | 0.224 | 0 | 0 | 0 | 0 | 8 |
| Num.of.lorry.pol | 0.002 | 0.068 | 0 | 0 | 0 | 0 | 4 |
| Num.of.trailer.pol | 0.011 | 0.116 | 0 | 0 | 0 | 0 | 3 |
| Num.of.tractor.pol | 0.034 | 0.250 | 0 | 0 | 0 | 0 | 6 |
| Num.of.agricultural.machines.pol | 0.005 | 0.110 | 0 | 0 | 0 | 0 | 6 |

Tabela 4 prikazuje prvih 10 atributa koji govore o broju različitih osiguranja. Što se ove grupe atributa tiče, primetno je da su standardne devijacije ponovo jako niske, ovoga puta sve ispod 1, dok se na prva dva mesta takođe nalaze osiguranje automobila i od požara. Samo osiguranje automobila ima korisnika sa 12 polisa, dok čak 4 tipa osiguranja nemaju ni je dnog korisnika sa brojem osiguranja 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 socio-demografski 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 klasifikaicije korisno je poslužiti se metodama redukcije atributa što će i biti tema u kasnijem delu rada.

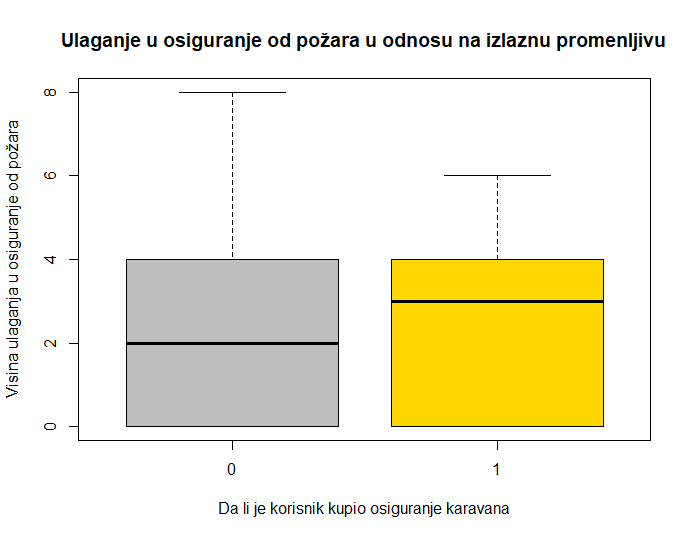
### Multivarijaciona eksploratorna analiza

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 automobila i osiguranja od požara, tako da će ovaj deo rada biti posvećen vizuelnom prikazu odnosa ova dva atributa sa izlaznom promenljivom.



SLika 2: Ulaganje u osiguranje automobila u odnosu na izlaznu promenljivu

Sa slike 2 može se primetiti velika razlika ponašanja korisnika u pogledu ulaganja u osiguranje automobile i kupovine osiguranja karavan vozila. Postoji samo jedan izuzetak u vidu korisnika koji nije osigurao automobil a jeste karavan vozilo. Zapaža se, takođe, velika razlika u medijani u ove dve grupe. Više od pola korisnika koji nisu kupili karavan osiguranje, nisu kupili ni osiguranje automobila, odnosno nisu ulagali u njega, dok polovina korisnika koji su osigurali karavan, uložili su preko hiljadu novčanih jedinica u osiguranje automobila. Ova raznolikost govori u prilog tome da će promenljiva koja se odnosi na ulaganje u osiguranje automobile definitivno imati značaja prilikom klasifikacije.



Slika 3: Ulaganje u osiguranje od požara u odnosu na izlaznu promenljivu

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 karavan 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.

## Redukcija atributa

Nakon stečenog uvida u određene karakteristike, odnosno paterne kod promenljivih u skupu podataka, sledeći korak predstavlja redukcija, odnosno smanjivanje broja promenljivih koje će klasifikator koristiti. Pitanje koje se nameće samo po sebi je zašto uopšte redukovati broj promenljivih i na taj način izgubiti određeni deo informacija koje pruža originalni skup podataka. Zapravo postoji više razloga koji su zaslužni za ovaj korak pripreme podataka.

Prvi je mogućnost nastajanja “overfitting-a”, problema vrlo čestog za visokodimenzionalne skupove podataka. Naime, ukoliko je broj atributa veliki, povećava se mogućnost da model previše dobro opiše trening podatke, ali da izgubi sposobnost generalizacije nad novim uzorcima. Kako je cilj mašinskog učenja što tačnije predviđanje, odnosno klasifikovanje, lako se uočava da previlika preciznost na trening setu zapravo nije ni blizu objektivan prikaz kvaliteta nekog modela.

Drugi razlog je taj što je želja svakog projektanta modela da on bude jednostavan i razumljiv, što se lako izgubi ukoliko je broj promenljivih veliki.

Sledeći razlog je takozvani “Garbage In Garbage Out” problem. Loš kvalitet ulaza proizvešće loš kvalitet izlaza. Ovo je čest slučaj ukoliko skup podataka sadrži veliki broj takozvanih neinformativnih atributa (<https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2>).

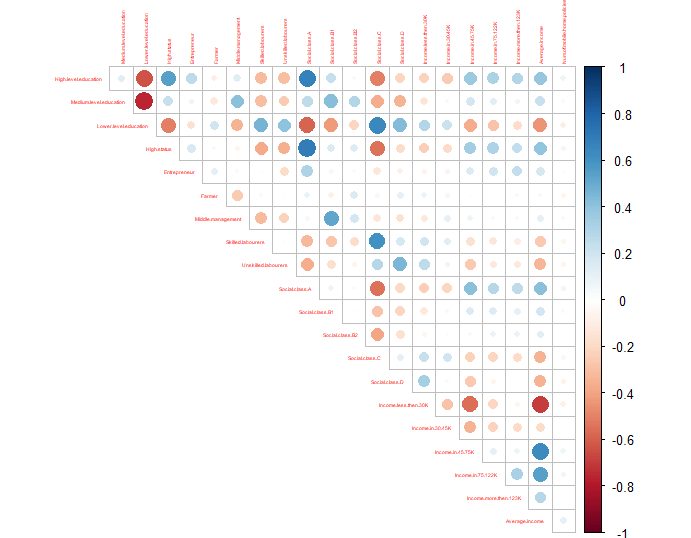
Kako skup podataka namenjen za ovaj projekat sadrži čak 86 atributa, od kojih smo videli da ne mali broj ne nosi značajnu informaciju, jasno je da je smanjenje broja promenljivih poželjno da ne bi došlo do problema nastalih zbog gorepomenutih slučajeva.

U praksi postoji mnoštvo metoda za redukciju atributa, međutim većina se može podeliti u tri grupe: filter metode, obmotavajuće metode (eng. wrapper-based) i ugrađene metode (eng. embedded). U ovom radu, prvo je korišćena jedna filter metoda, na koju je nadovezana jedna obmotavajuća metoda da bi se dobio konačan skup podataka spreman za upotrebu od strane modela (<https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2>).

### Filter metode

Ovaj tip metoda zasniva se na rangiranju svakog atributa jednom od univarijacionih metrika i potom biranju atributa sa najvišim rangom. Neke od univarijacionih metrika su varijansa, “chi-square”, koeficijenti korelacije, kao i informaciona dobit (<https://towardsdatascience.com/feature-selection-in-python-using-filter-method-7ae5cbc4ee05>).

U ovom radu korišćena je metrika koeficijenta korelacije. Kako se skup podataka sastoji od ordinalnih atributa, a uslovi za korišćenje pirsonovog koeficijenta korelacije nisu zadovoljeni, odlučeno je da se koristi spirmanov koeficijent korelacije (<https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php>). Ispitana je korelacija između određenih socio-demografskih karakteristika za koje se smatralo da bi mogle biti visoko korelisane, što se ispostavilo kao slučaj sa promenljivima koje se tiču obrazovanja, visine prihoda i pripadnosti određenoj socijalnoj klasi. Nakon ovog ispitivanja odlučeno je da se izbace promenljive koje se tiču pripadnosti socijalnoj klasi jer su one bile u najslabijoj korelaciji sa izlaznom promenljivom u odnosu na ostale pomenute. Iz istog razloga neće se koristiti promenljive koje se odnose na edukaciju. Takođe, iz grupe sociodemografskih atributa izbacuju se oni koji se odnose na religiju koja nema praktičnog uticaja na to da li će osoba kupiti karavan osiguranje. Takođe, informacije o prihodu su sadržane u promenljivim koje obuhvataju date opsege prihoda, tako da atribut koji govori o prosečnom prihodu može da se izbaci iz skupa podataka.



Slika 4: Tabela korelacije

Kao što je već pokazano u prethodnom delu rada, kategorije promenljivih koje se odnose na ulaganje novca u određene tipove osiguranja i broj kupljenih osiguranja nose malu informaciju pre svega jer su to vrste osiguranja koja se ne susreću toliko često u praksi, pa samim tim većina korisnika ih i ne kupuje, odakle proizilazi da je veliki broj polja u skupu podataka jednak nuli. Malo detaljnijom analizom dolazi se do zaključka da samo tri vrste osiguranja imaju broj polja jednak nuli koji iznosi manje od 90% i kada je u pitanju ulaganje novca i kada je u pitanju broj kupljenih osiguranja. Treba napomenuti da sva tri tipa, a to su osiguranje automobila, osiguranje od požara i osiguranje treće strane imaju značajno više korisnika od svih ostalih osiguranja tako da se za sada ostavljaju u skupu podataka koji će se koristiti za klasifikaciju. Pored njih u ovom skupu ostaće i osiguranje broda s obzirom na visoku korelaciju sa izlaznom promenljivom. Sve ostale vrste osiguranja nose vrlo malu informaciju tako da će se agregirati jednostavnim sabiranjem i na taj način dobijaju se još dve promenljive, jedna koja se odnosi na ulaganja i druga koja se odnosi na broj kupljenih osiguranja, koje ulaze u skup podataka.

### Obmotavajuće metode – Boruta algoritam

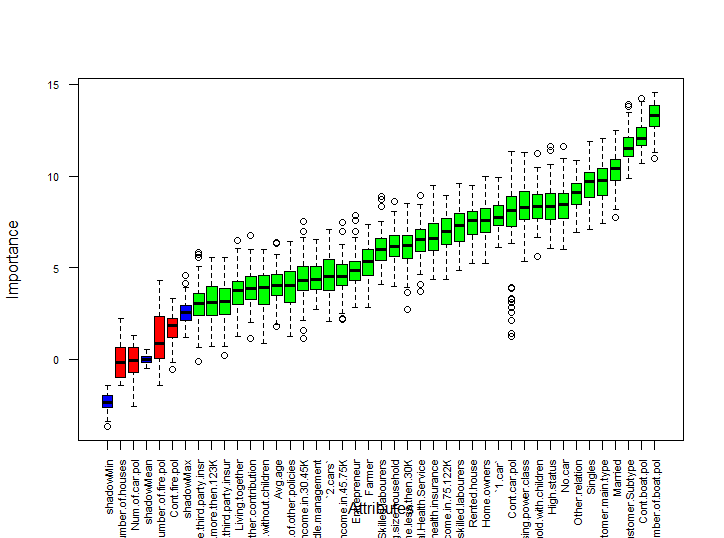
U prethodnom delu, pomoću filter metoda redukcije podataka i iskustva, odnosno logičkog zaključivanja, originalan skup podataka smanjen je sa 86 atributa na 41. Taj broj je pogodniji za upotrebu u modelima, jer u određenoj meri umanjuje šanse za nastanak problema u slučaju velikog broja promenljivih koji su opisani u prethodnom delu rada. Međutim, postoji li način na koji se može proveriti da li svi preostali atributi stvarno nose značajnu informaciju, odnosno da li će imati uticaja prilikom klasifikacije? Naravno, moguće je sprovesti još neku metodu redukcije podataka. U ovom delu rada biće opisana jedna veoma korisna tehnika, koja spada u obmotavajuću grupu metoda i to oko algoritma slučajne šume, a naziva se Boruta algoritam.

Boruta algoritam sastoji se iz nekoliko koraka:

1. Prvo, dodaje određenu dozu nasumičnosti postojećem skupu podataka tako što kreira kopije svih promenljivih koje se sastoje od nasumično izmešanih vrednosti originalnih promenljivih. Ovi atributi se nazivaju „shadow“ atributi.
2. Nakon toga, boruta trenira klasifikator slučajne šume na proširenom skupu podataka i primenjuje meru značajnosti (podrazumevana je “mean decrease accuracy”) da bi se evaluirao značaj svakog atributa.
3. U svakoj iteraciji, ovaj algoritam proverava da li originalni atribut ima veći značaj od najboljeg od svih „shadow“ atributa (da li originalna promenljiva ima veći Z skor od maksimalnog Z skora među odgovarajućim „shadow“ atributima) i izbacuje atribute koji se pokazuju neznačajnim.
4. Na kraju, algoritam prestaje sa radom ili kada se sve promenljive pokažu značajnim, odnosno neznačajnim, ili kada se dostigne određeni limit krugova slučajne šume.

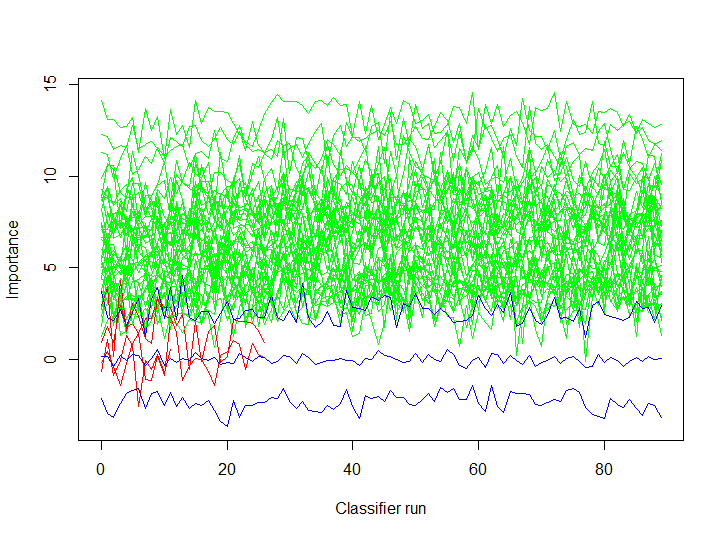
(<https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/>)

Nakon 90 iteracija algoritam je pronašao čak 36 značajnih promenljivih i samo 4 neznačajne što ide u prilog tome da je prvobitna redukcija podataka odrađena na dobar način. Među neznačajnim atributima nalazi se samo 1 koji pripada grupi socio-demografskih karakteristika i to je u pitanju broj kuća. Ostale tri, ulaganje u osiguranje od požara i broj osiguranja od požara, kao i broj osiguranja automobila, iako imaju značajno više pojavljivanja od drugih atributa koji se odnose na vrste osiguranja, takođe su proglašene neznačajnim. Sve ovo ide u prilog činjenici da će glavnu ulogu u određivanju vrednosti izlazne promenljive imati socio-demografske karakteristike.



Slika 5: Grafička reprezentacija boruta algoritma

Na slici su prikazani boks-plotovi za svaki atribut kada je u pitanju njegov značaj za klasifikaciju izlazne promenljive, kao i boks-plotovi minimalnih, srednjih i maksimalnih vrednosti „shadow“ atributa. Idealno, „shadow“ atributi ne bi trebalo da imaju veliki značaj. Zapravo, značaj ovih atributa treba da bude blizak nuli, ali zbog nasumične fluktuacije može imati vrednost različitu od nule. Dakle, značaj „shadow“ atributa može se koristiti kao referentna vrednost za određivanje koliko je značajan svaki atribut. Na slici su s leva na desno poređani atributi po značaju u rastućem redosledu. Vidimo na početku prvo 4 crvena boks-plota koji označavaju 4 pomenute promenljive koje nisu prošle test značajnosti, dok posle njih slede sve one promenljive koje su prošle test.



Slika 6: Grafička reprezentacija kretanja značajnosti atributa kroz iteracije

Moguće je iscrtati i istoriju značaja atributa. Ako vrednost pada u okviru plavih linija, te promenljive najverovatnije neće proći test značajnosti, dok u zelenu oblast upadaju promenljive koje imaju mnogo veći značaj nego „shadow“ atributi.

Tabela 6: Značaj atributa tokom iteracija

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | meanImp | medianImp | minImp | maxImp | normHits | decision |
| Cont.private.third.party.insr | 3.015 | 3.054 | -0.113 | 5.821 | 0.711 | Confirmed |
| Cont.car.pol | 7.672 | 8.136 | 1.248 | 11.362 | 0.944 | Confirmed |
| Cont.fire.pol | 1.667 | 1.867 | -0.522 | 3.308 | 0.044 | Rejected |
| Cont.boat.pol | 12.129 | 12.058 | 10.660 | 14.203 | 1.000 | Confirmed |
| Num.of.private.third.party.insr | 3.171 | 3.165 | 0.195 | 5.541 | 0.711 | Confirmed |
| Num.of.car.pol | -0.132 | -0.052 | -2.580 | 1.296 | 0.000 | Rejected |
| Num.of.fire.pol | 1.089 | 0.845 | -1.433 | 4.298 | 0.044 | Rejected |
| Num.of.boat.pol | 13.164 | 13.294 | 10.947 | 14.575 | 1.000 | Confirmed |
| Other.contribution | 3.929 | 3.888 | 1.166 | 6.733 | 0.878 | Confirmed |
| Num.of.other.pol | 4.047 | 4.035 | 1.230 | 6.425 | 0.889 | Confirmed |

Na osnovu tabele 6 može se analizirati dodatna statistika koja govori kako se kretao značaj atributa tokom iteracija. Data statistika pokazuje koliko su iznosile minimalna, maksimalna, srednja vrednost kao i medijana vrednosti značaja svakog atributa. Pored toga, “normHits” govori koji procenat vremena se svaka promenljiva pokazala značajnijom nego „shadow“ atribut. Vidimo da je ovaj procenat blizak ili jednak nuli za attribute koji nisu prošli test, dok je za ostale daleko iznad 50 %, a za mnoge je i jednak 100 %, što govori da bi ti atributi trebalo da imaju veliku ulogu u klasifikatoru.

Nakon redukcije atributa, originalni skup podataka sveden je na završni koji se sastoji od 36 promenljivih za koje se smatra da će imati najveći značaj prilikom klasifikovanja izlazne promenljive, tako da su se stekli svi preduslovi za izvođenje eksperimenta.

# Prilog: Kod

## Struktura podataka

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

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

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

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

summary(train.data)

str(train.data)

table(train.data$Num.of.mobile.home.policies)

plot(table(train.data$Num.of.mobile.home.policies), type = "h", col = "red", lwd = 15, xlab = "Da li je korisnik kupio osiguranje: 0 - Ne, 1 - Da", ylab = "Broj korisnika", main = "Prikaz nebalansirane strukture izlazne promenljive")

## Univarijaciona analiza podataka

str(train.data)

summary(train.data)

# prva kategorija

apply(data[6:43], 2, summary)

apply(data[6:43], 2, sd)

# druga kategorija

uni.eda.cont <- apply(data[44:64], 2, summary)

apply(data[44:64], 2, sd)

class(uni.eda.cont)

data.uni.eda.cont <- as.data.frame(uni.eda.cont)

sum(data.uni.eda.cont["Max.",]==9)

sum(data.uni.eda.cont["Median",]!=0)

sum(data.uni.eda.cont["3rd Qu.",]!=0)

# treća kategorija

uni.eda.num.of <- apply(data[65:85], 2, summary)

apply(data[65:85], 2, sd)

class(uni.eda.num.of)

data.uni.eda.num.of <- as.data.frame(uni.eda.num.of)

sum(data.uni.eda.num.of["Max.",]==12)

sum(data.uni.eda.num.of["Max.",]==1)

sum(data.uni.eda.num.of["Median",]!=0)

sum(data.uni.eda.num.of["3rd Qu.",]!=0)

## Multivarijaciona analiza podataka

by(data$Cont.car.pol, data$Num.of.mobile.home.policies, summary)

by(data$Cont.fire.pol, data$Num.of.mobile.home.policies, summary)

boxplot(data$Cont.car.pol~data$Num.of.mobile.home.policies, col = c("grey","gold"),

main = "Ulaganje u osiguranje automobila u odnosu na izlaznu promenljivu",

xlab = "Da li je korisnik kupio osiguranje karavana",

ylab = "Visina ulaganja u osiguranje automobila")

boxplot(data$Cont.fire.pol~data$Num.of.mobile.home.policies,

col = c("grey","gold"), main = "Ulaganje u osiguranje od požara u odnosu na izlaznu promenljivu",

xlab = "Da li je korisnik kupio osiguranje karavana",

ylab = "Visina ulaganja u osiguranje od požara")

## Filter metode

install.packages("corrplot")

library(corrplot)

train.data[,86] <- as.integer(train.data[,86])

corr\_matrix<- cor(train.data[,c(16:29,37:42,86)])

corr\_matrix<- cor(train.data[,c(16:29,37:42,86)], method = "spearman")

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)

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)

corr\_matrix\_3<- cor(data[,c(16:18,25:29,37:41,86)])

corrplot(corr\_matrix\_3, type = "upper", diag = F,tl.pos = NULL, tl.cex = 0.3, tl.col = "red", tl.offset = 0.4, tl.srt = 90)

#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

## Obmotavajuće metode – Boruta algoritam

# Boruta algoritam

library(Boruta)

set.seed(111)

borutaAlgorithm <- Boruta(Num.of.mobile.home.policies ~ ., data = train.data.new, doTrace = 2)

print(borutaAlgorithm)

borutaAlgorithm$finalDecision

# Varijable koje nisu odbacene

nonRejected <- getNonRejectedFormula(boruta)

# Varijable koje su potvrdjene kao znacajne

confirmed <- getConfirmedFormula(boruta)

# las = 2 - nazivi atributa da bufduvertikalni, cex.axis - velicina fonta

plot(borutaAlgorithm, las = 2, cex.axis = 0.7)

plotImpHistory(borutaAlgorithm)

attStats(borutaAlgorithm)