Tveganje finančnega zakupa (lizinga) avtomobilov

Jan Lampič, Fakulteta za matematiko in fiziko

21. avgust 2018

Sem Jan Lampič, študent 3.letnika finančne matematike. V tem poročilu vam bom na kratko predstavil izgradnjo ter rezultate modela, ki napoveduje če bo za predmet lizinga (v našem primeru avtomobila) sklenjena pogodba. Seminarska nalogo sem naredil v okviru predmeta Izbrane teme iz analize podatkov (ITAP). Pri izgradnji modela sem uporabljal programski jezik `R`.

1 Podatkovna množica

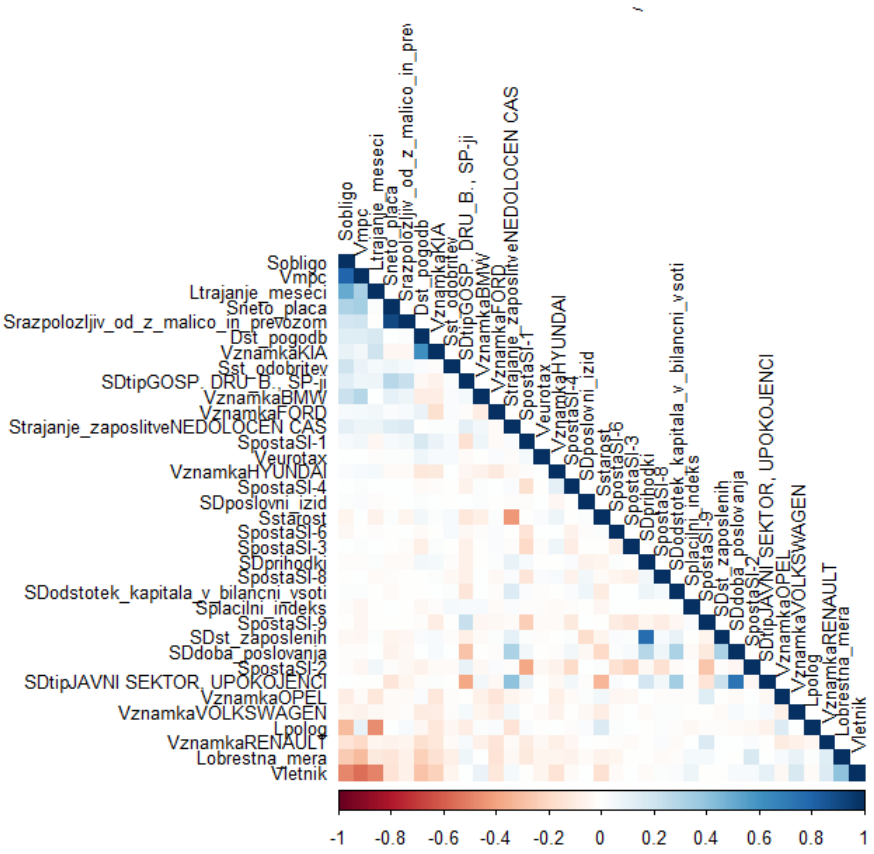
Učna podatkovna množica vsebuje 6.062 predmetov zakupa (leasinga) vozil, ki jih je leta 2015 obravnavalo izbrano leasing podjetje v Sloveniji. Vsak predmet je opisan z enaintridesetimi (31) spremenljivkami, ki se nanašajo na lastnosti stranke (S), njenega delodajalca (SD), leasing predmeta (L), vozila (V) in njegovega dobavitelja (D). Ciljna spremenljivka je *Lodobren*, ki pove ali je pogodba sklenjena ali ne.

2 Uvoz in transformacija podatkovne množice

Začel sem z uvozom učne in testne podatkovne množice. Nato sem uredil obe podatkovni množici. Vse stolpce (spremenljivke) sem pretvoril v ustrezen tip (`factor` in `numeric`) in normaliziral učno ter testno množico. Pri pregledu množice sem opazil, da imata stolpca *Splacilni_indeks* in *Szapadlo_neplacano* manjkajoče podatke. Te sem nadomestil z metodo `na.roughfix` iz knjižnice `randomForest`, ki nadomesti manjkajoče vrednosti s povprečjem oz. mediano znanih vrednosti.

3 Izbira pojasnevalnih spremenljivk

Za izbiro pojasnevalnih spremenljivk sem moral vse diskretne spremenljivke pretvoriti v naivne (angl. dummy variables) zato, da sem lahko izračunal variance in kovariance spremenljivk. Potem sem z metodo `nearZeroVar` iz knjižnice `caret` odstranil vse spremenljivke z nizko varianco. Pri tem je bilo odstranjenih 68 spremenljivk. Nato sem izračunal še kovariančno matriko in odstranil visoko korelirane spremenljivke (tiste, ki so imele korelacijo nad 0.85). Koreliranost med podatki prikazuje graf 1.



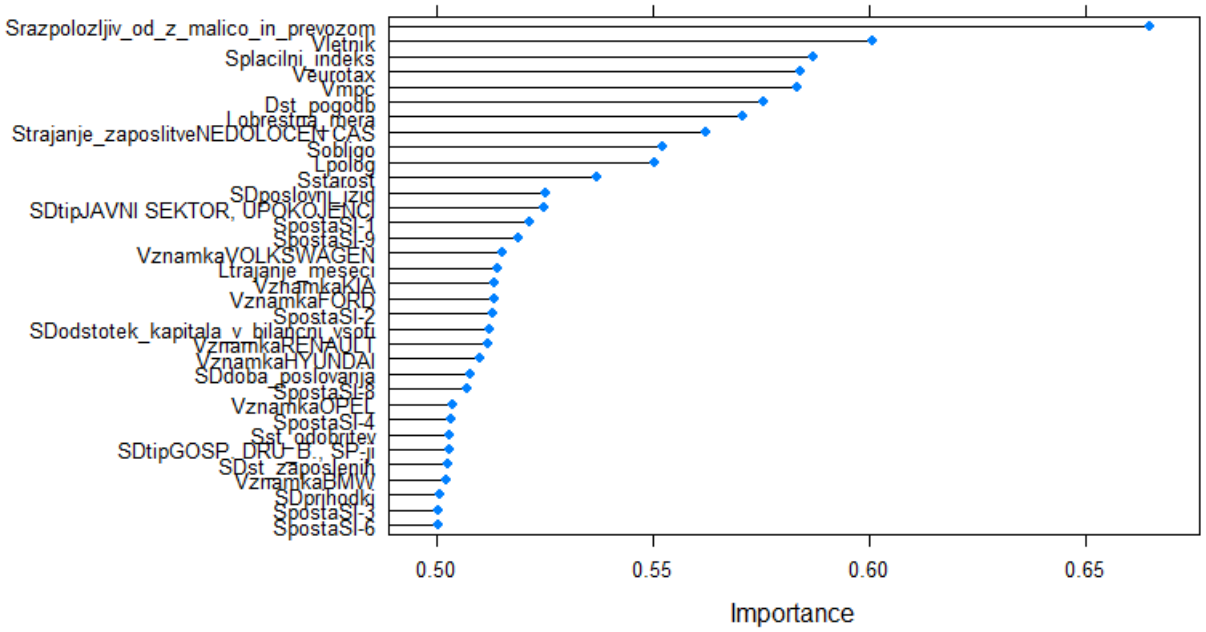
Slika 1: Graf koreliranosti med spremenljivkami

Pri izbiri spremenljivk pa sem si pomagal z izgradnjo LVQ modela (Learning Vector Quantization), ki oceni pomembnost spremenljivk. Tako sem dobil predstavo katere spremenljike so pomembne (glej graf 2). Nato pa sem še z metodo RFE (Recursive Feature Elimination) knjižnice `caret` izbral najbolj pomembne spremenljivke. Metoda z algoritmom Random Forest na vsaki iteraciji oceni model, pri tem pa razišče vse možne podmnožice pojasnevalnih spremenljivk. Iz grafa 3 je razvidno, da je model najbolj natančen pri uporabi 16 pojasnevalnih spremenljivk, to so:

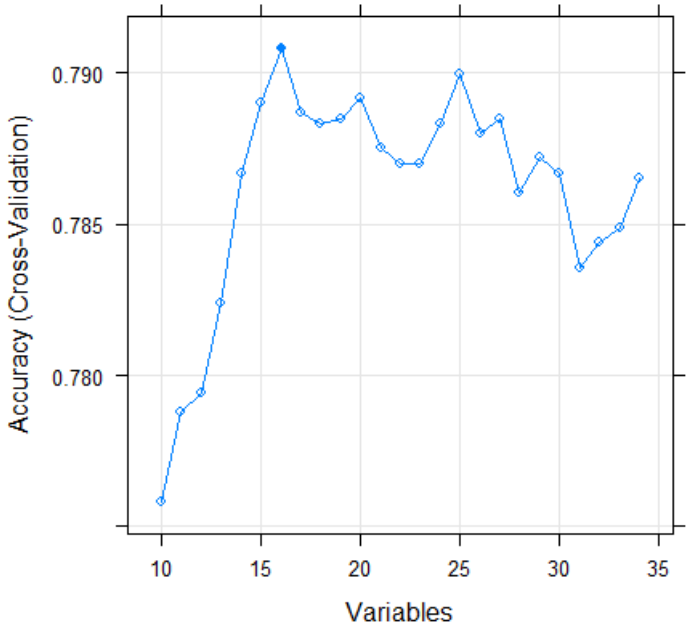
- $Srazpolozljiv\_od\_z\_malico\_in\_prevozom$,
- $Splaciln\_indeks$,
- $Veurotax$,
- $Strajanje\_zaposlitveNEDOLOCENCAS$,
- $Sstarost$,
- $Lpolog$,

- $Vletnik$,
- $SDtipJAVNISEKTOR,UPOKOJENCI$,
- $Vmpc$,
- Dst_pogodb,
- $Sobligo$,
- $SDodstotek\_kapitala\_v\_bilancni\_vsoti$,
- $SDposlovni\_izid$,
- $SDst\_zaposlenih$,
- $SDprihodki$,
- $Lobrestna\_mera$.

Na koncu pa sem popravil še razmerje ciljne spremenljivke, saj je bila porazdeljena neenakomerno, ker je bilo več kot 60% odobrenih pogodb. Z metodo `SMOTE` (kombinacija podvzorčenja in prevzorčenja) sem razmerje sklenjenih pogodb proti nesklenjenim pretvoril na 47:53. Tako je imela učna množica 16 napovednih spremenljivk in 5953 primerov.



Slika 2: Graf pomembnosti spremenljivk



Slika 3: Ocenjena natančnost modela pri različnem številu izbranih spremenljivk

4 Učenje modela

Pri izgradnji modela sem najprej definiral napovedno napako, kjer sem minimiziral $\frac{1}{2}\frac{FN}{TP+FN} + \frac{FP}{TN+FP}$. Izrazu $\frac{FN}{TP+FN}$ sem dal manjšo utež, saj domnevam, da so stroški pri nesklenjeni dobri pogodbi manjši kot pri sklenjeni slabi pogodbi. Napovedno napako sem meril z 10-kratnim prečnim preverjanjem.

Za izbiro najboljšega modela sem definiral funkcijo `bestModel`. Funkcija prejme učno množico nato pa na njej nauči šest modelov pri različnih parametrih in vrne tistega z najmanjšo napovedno napako. Funkcija primerja med naslednjimi modeli:

- metodo najbližjih sosedov,
- logistično regresijo,
- odločitvenim drevesom,
- linearno metodo podpornih vektorjev,
- metodo naključni gozdovi,
- metodo podpornih vektorjev z radialnim jedrom.

Za najboljši model se je izkazal model naključnih gozdov (random forest), s katerim sem potem napovedal vrednosti testne množice. Napovedi modela so shranjene v datoteki `napovedi.csv`, skripta za ureditev podatkov in izgradnjo modela pa je dostopna na mojem GitHub repozitoriju (`https://github.com/LampicJ15/ITAP`).