Fakultet organizacionih nauka, Beograd – Otkrivanje zakonitosti u podacima

**„Analiza marketinške kampanje banke za otvaranje oročenog depozita“**

**Damir Pajaziti,**

19. Mart 2020.

# Uvod

„Ukoliko bih imao određenu svotu novca u svom džepu kao celokupan budžet svoje kompanije, 90% bih potrošio na marketing, a ostatak na samu uslugu ili proizvod“. Ovakav citat, i njemu slični, mogu se neretko pronaći na internetu kao putokaz do uspeha mnogih start-up kompanija. Marketing je proces u kome kompanije kreiraju vrednost za svoje korisnike i uz pomoć kojeg grade jake veze sa korisnicima u cilju dobijanja vrednost od strane korisnika za uzvrat. Termin „oročeni depozit“ je ulaganje sa unapred dogovorenom fiksnom kamatnom stopom. Nakon isteka oročenja klijent može podići svoj novac ili dogovoriti novi rok i produžiti period oročenja. Banke kao institucije imaju poslovni interes da što više klijenata oroči svoj novac po određenoj kamatnoj stopi, kako bi ga prebacile u druge novčane tokove odnosno nastavile poslovanje ali u okviru drugih modela poslovanja kao što su npr. krediti, osiguranja itd.

# Podaci

U ovom radu koriste se podaci koji su javno dostupni za istraživanje [Moro at al., 2014] [[1]](#footnote-1) [[2]](#footnote-2). Podaci su obogaćeni dodatkom 5 novih socijalnih i ekonomskih karakteristika, objavljenih od strane portugalske banke „Banco de Portugal“, i javno su dostupni[[3]](#footnote-3). Korišćeni set podataka je skraćena verzija koja sadrži samo 10% nasumično odabranih zapisa od verzije fajla koja je kompletna. Skraćena sadrži ukupno 4119 zapisa. Prednost korišćenja manje verzije je u brzini testiranja algoritama koji imaju veći broj kalkulacija. Podaci se nalaze u fajlu koji je u csv formatu. U setu se nalazi ukupno 20 nezavisnih atributa i jedan zavisan. U nekim zapisima nalaze se nedostajući podaci koji su predstavljeni sa NaN ili u slučaju kategoričkih varijabli „unknown“ vrednostima. U setu podataka imamo 10 numerickih nezavisnih kolona (*age, duration, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed*) i 10 kategoričkih kolona (*job, marital, education, default, housing, loan, contact, month, day\_of\_week, poutcome*).

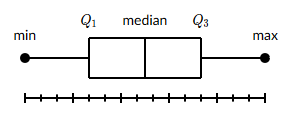
# Priprema podataka za prediktivno modelovanje

Prema procenama analitičara, priprema podataka za prediktivno modelovnje je mukotrpan proces koji čini oko 50% ukupnog rada. U ovom radu, primenjeno je više različitih tehnika za: umetanje nedostajućih podataka, pronalaženje ekstremnih vrednosti, pretvaranje kategoričkih u numeričke podatke, redukciju kolona itd. Za umetanje nedostajućih kategoričkih podataka korišćene su tehnike *umetanje najčešćih vrednosti* (pomoću funkcije *mode()*), umetanje prosečnih vrednosti (pomoču funkcije *mean*()), dok su za imputaciju kategoričkih promenljivi korišćena imputacija najčešćih vrednosti (pomoću funkcije mode()), prediktivni modeli (linearna regresija), a u nekim slučajevima nedostajuće vrednosti su smatrane novom kategorijom. Svaka od ovih tehnika ima svoje prednosti i mane, u pogledu na efikasnost, brzinu izvršavanja, promene varijanse u skupu podataka itd.

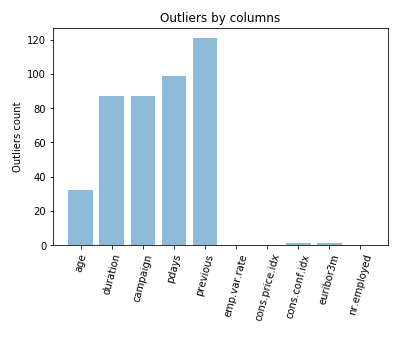
Za detekciju ekstremnih vrednosti korišćena su dva pristupa.

Zscore:

i box-plot:



Primećeno je da kolone *age, duration, pdays, campaign, previous* imaju dosta vrednosti koje su veće od 3z, ali logičkiom proverom raspona vrednosti nismo uspeli da sa sigurnošću utvrdimo da li su to izuzeci ili greške u podacima. Čist primer je kolona age koja ima tridesetak zapisa preko 3z ali je maksimalna vrednost 80 godina pa te vrednosti ne možemo smatrati neispravnim. Takođe je pronađen jedan zapis koji zaista ima ekstremne vrednosti (u nekim kolonama i preko 65 standardnih devijacija udaljen od proseka) pa je i uklonjen.



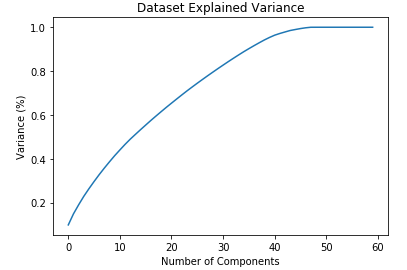
Konvertovanjem kategoričkih u numeričke podatke dobijamo veći raspon algoritama koje možemo koristiti za prediktivnu analizu. Svi algoritmi koji koriste metrike za distrance (npr. euklidsko rastojanje)

U radu je korišćena strategija *dummy coding*[[4]](#footnote-4) za pretvaranje kategoričkih u numeričke podatke. Ovoj tehnici data je prednost u odnosu na tehniku *label encoder* [[5]](#footnote-5)kako bismo izbegli probleme sa regresionim algoritmima koji su osetljivi na sortirani brojeve (npr. 5 raličitih radnih mesta bi bila enkodirana u niz od 5 brojeva 1,2,3,4,5) pa bi se mogla zaključiti pogrešna zakonitost u podacima vezana za veličinu broja.

# Analiza glavnih komponenti

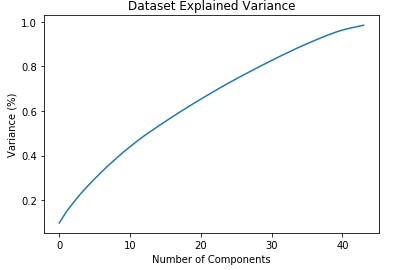
Analiza glavnih komponenti (eng. *Principal Component Analysis*) u daljem tekstu PCA, je jedna od najkorišćenijih tehnika za redukciju podataka. Njen cilj je da smanji količinu podataka, a da pritom ti podaci nose približnu količinu korisnih informacija. Smanjivanjem 1% ili 2% varijanse u setu podataka možemo redukovati jako puno nepotrebnih podataka (komponenti) u našem setu. PCA radi tako što transformiše izvorne podatke i njihove dimenzije u nove, nama teško razumljive. Pre PCA transformacije potrebno je skalirati podatke. U ovom radu je korišćena sledeća formula:

Prethodne dimenzije podataka (kolone) nakon PCA transformacije postaju komponente. Prva komponenta ima najveću varijansu, a svaka sledeća ima manju od prethodne. Kumulativno zajedno imaju 100% varijanse i isti je broj komponenata kao i originalnih kolona (*slika 1*).



Slika – broj komponenti sa 100% kumulativne varijanse

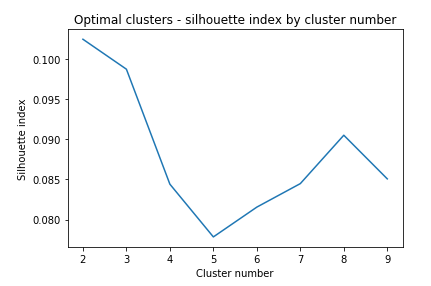
Nakon PCA transformacije broj komponenata je predstavljen na *slici 2*:



Slika – broj komponenti sa 98% kumulativne varijanse

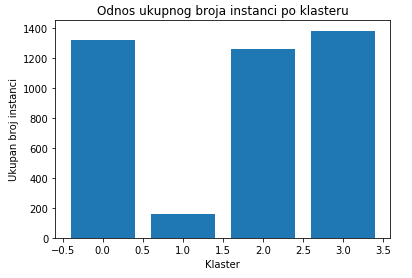
# Klasterovanje

U ovom radu za izvršenje zadatka klasterovanja korišćen je *K-Means* algoritam odnosno njegova unapređena verzija *K-means++,* koja ima poboljšane korake inicijalizacije. Ovaj algoritam klasterovanja ima pored inicijalizacije još nekoliko slabih tačaka. Korisnik mora sam odrediti unapred broj klastera. Da bi smo prevazišli ovaj problem, korišćena mera evaluacije klastera *Silhouette Index Score* kako bismo odredili optimalan broj klastera (*slika 3*)*.*



Slika – Silhouette index score by cluster number

U samom zahtevu zadatka traženo je da se podaci podele na onoliki broj klastera koliki je broj prediktivnih modela koji će kasnije biti korišćeni. U ovom radu korišćeno je 4 modela, pa je u skladu sa time set podataka podeljen na 4 klastera (*slika 4*).



Slika – prikaz ukupnog broja instance po klasteru

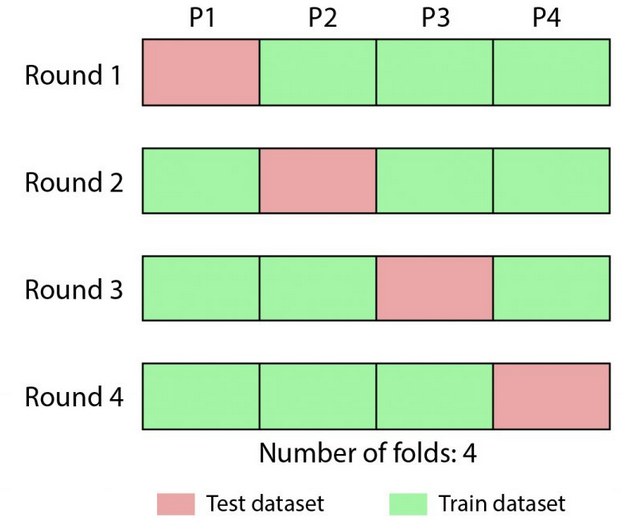
# Prediktivni modeli i njihova evaluacija

Za izdradu prediktivnih modela korišćeni su sledeći algoritmi mašinskog učenja:

* Naïve Bayes
* Logistička regresija
* KNN
* Decision three
* Random Forest

Pre treniranja prediktivnih modela, podaci su normalizovani *MinMaxScaler()*

ili sa već pomenutim *StandardScore()* algoritmom. Kreirani su pajpovi (*eng. Pipeline*) i pušteni algoritam za evaluaciju cross\_validate[[6]](#footnote-6) (*slika 5*).



Slika – cross validation ilustrated

Rezultati evaluacije logističke regresije:

fit\_time 0.111272

score\_time 0.006598

test\_precision 0.433547

train\_precision 0.450673

test\_recall 0.813118

train\_recall 0.863750

test\_accuracy 0.867109

train\_accuracy 0.874123

test\_f1 0.564533

train\_f1 0.592279

Rezultati evaluacije naivnog bajesa:

fit\_time 0.038092

score\_time 0.006941

test\_precision 0.397580

train\_precision 0.396430

test\_recall 0.554516

train\_recall 0.551551

test\_accuracy 0.861550

train\_accuracy 0.863636

test\_f1 0.460456

train\_f1 0.461246

Rezultati evaluacije random forest ansambla:

fit\_time 0.488958

score\_time 0.023283

test\_precision 0.684345

train\_precision 1.000000

test\_recall 0.285376

train\_recall 1.000000

test\_accuracy 0.908737

train\_accuracy 1.000000

test\_f1 0.397461

train\_f1 1.000000

Rezltati evaluacije knn algoritma:

fit\_time 0.081918

score\_time 0.158324

test\_precision 0.535952

train\_precision 0.827447

test\_recall 0.253118

train\_recall 0.450646

test\_accuracy 0.897290

train\_accuracy 0.931915

test\_f1 0.336688

train\_f1 0.583380

Rezultati evaluacije decision three algoritma:

fit\_time 0.031869

score\_time 0.007958

test\_precision 0.485377

train\_precision 1.000000

test\_recall 0.436129

train\_recall 1.000000

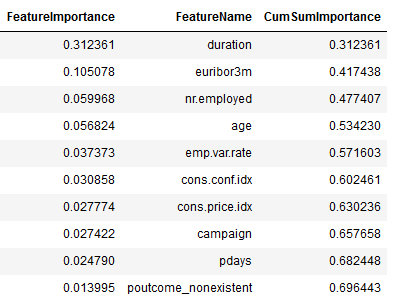
test\_accuracy 0.890006

train\_accuracy 1.000000

test\_f1 0.452689

train\_f1 1.000000

Jedan od glavnih zadataka je pronaći optimalan algoritam koji daje najbolje rezultate prilikom evaluacije. U ovom radu najbolji score je imao Random Forest algoritam. Kako bi povećali tačnost algoritma, koristili smo tehniku *over sampling[[7]](#footnote-7)*, kako bismo dobili balansirane izlazne klase. Sledeći izazov je izabrati prave parametre algoritama i za tu svrhu korišćena je tehnika *GridSearch[[8]](#footnote-8).* Nakon toga tražene su kolone koje imaju najveći značaj (*eng. Feature importance*) pomoću metode *SelectKBest*[[9]](#footnote-9) *(slika 6)*.



Slika – važnosti prvih 10 atributa u setu podataka

Nakon prethodnih optimizacija I prilagođavanja, pokrenut je ponovo algoritam *Random Forest* koji je dao sledeće rezultate:

fit\_time 0.264135

score\_time 0.012851

test\_precision 0.977644

train\_precision 1.000000

test\_recall 0.881341

train\_recall 0.996486

test\_accuracy 0.933321

train\_accuracy 0.998243

test\_f1 0.888558

train\_f1 0.998239

# Zaključak

Prilikom kreiranja prediktivnih modela potrebno je posebno obratiti pažnju na selekciju atributa, i pronalaženje parametara za algoritme. Priprema podataka je mukotrpan proces koji je preduslov za rad mnogih algoritama. U ovom radu nedostajuće vrednosti su umetnute pa nisu brisani zapisi iz seta podataka. Potrebno je obratiti pažnju na balans izlazne klase a kao preporuku za evaluaciju koristiti F1 skor. Na taj način možemo biti sigurni da svaka izlazna klasa ima približan skor. F1 skor je dobar za nebalansirane podatke jer koristi harmonični prosek metrika.

# Budući rad

Različiti algoritmi ansambli su se u praksi pokazali kao dobri prediktori, i na takmičenjima često odnose visoko rangirane pozicije. Ideja je pokušati pronaći novi algoritam koji bi imao slične ili bolje ocene od Random Forest-a.

Kolona duration ima veliku korelaciju, i sam kreator seta podataka daje preporuku da se u produkcionom okruženju ta kolona ne koristi za predviđanje.

Pronaći nove kolone na internetu kako bi dobili još veću prediktivnu moć.

Evaluirati neki od neuronskih klasifikatora i uporediti rezultate sa ostalim algoritmima.

1. [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001 [↑](#footnote-ref-1)
2. <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing> [↑](#footnote-ref-2)
3. <https://www.bportugal.pt/estatisticasweb> [↑](#footnote-ref-3)
4. Dummy codding – pretvaranje kategoričkih u numeričke podatke kreiranjem nove kolone koja je tipa bit (oznaka 0 ili 1 govori da li je vrednost prisutna) [↑](#footnote-ref-4)
5. Label encoder – tehnika koja konvertuje kategoričke podake u sortirane brojeve. [↑](#footnote-ref-5)
6. Proces podele test i trening podataka na segmente, gde se samo jedan segment koristi kao test a ostali za treniranje. Korišćenjem ove tehnike evaluacije postiže se veća generalizacija modela. [↑](#footnote-ref-6)
7. Tehnika kreiranja novih zapisa u cilju balansiranja izlazne klase. [↑](#footnote-ref-7)
8. Tehnika pretrage optimalni parametara koja radi po metodi probaj sve i vrati najbolji. [↑](#footnote-ref-8)
9. Metoda koja selektuje atribute koji imaju najbolji skor. [↑](#footnote-ref-9)