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“, (Bill Gates, 1995.). Ovaj 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 da dobiju 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 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 NaN ili u slučaju kategoričkih varijabli „unknown“. 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*).

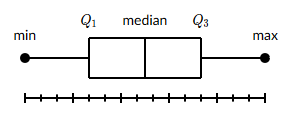
# Priprema podataka za prediktivno modelovanje

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.

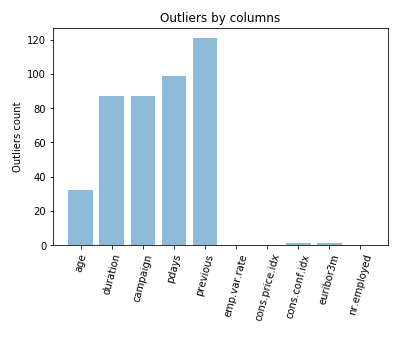
Za detekciju ekstremnih vrednosti korišćena su dva pristupa.

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

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)