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.

# Deskriptivna statistika

U ovom projektu korišćeni se podaci koji su javno dostupni za istraživanje [Moro at al., 2014] [[1]](#footnote-1) [[2]](#footnote-2). 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*). U prilogu je tabela sa prikazom statistike za numeričke kolone:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **age** | **duration** | **campaign** | **pdays** | **previous** | **emp.var.rate** | **cons.price.idx** | **cons.conf.idx** | **euribor3m** | **nr.employed** |
| **count** | 4119 | 4116 | 4114 | 4119 | 4117 | 4116 | 4117 | 4117 | 4117 | 4118 |
| **mean** | 40.1 | 256.87 | 2.5343 | -0.733 | 0.1904 | 0.084694 | 93.57954 | -40.4864 | 4.85899 | 5166.498 |
| **std** | 10.3 | 254.76 | 2.5666 | 1.53281 | 0.54190 | 1.563361 | 0.579442 | 4.63984 | 79.43246 | 73.66936 |
| **min** | 18 | 0 | 1 | -1 | 0 | -3.4 | 92.201 | -50.8 | 0.635 | 4963.6 |
| **25%** | 32 | 103 | 1 | -1 | 0 | -1.8 | 93.075 | -42.7 | 1.334 | 5099.1 |
| **50%** | 38 | 181 | 2 | -1 | 0 | 1.1 | 93.749 | -41.8 | 4.857 | 5191 |
| **75%** | 47 | 317.25 | 3 | -1 | 0 | 1.4 | 93.994 | -36.4 | 4.961 | 5228.1 |
| **max** | 88 | 3643 | 35 | 21 | 6 | 1.4 | 94.767 | 1.266 | 5099.1 | 5228.1 |

Izračunat je balans izlazne varijable 89.0265% za klasu sa vrednošću 0 (no), i 10.9493% za klasu 1 (yes), što nam govori da su podaci u zavisnoj varijable nebalansirani. Primećen je jedan zapis koji je NaN i on je zamenjen sa najčešćom vrednošću (funkcija mode).

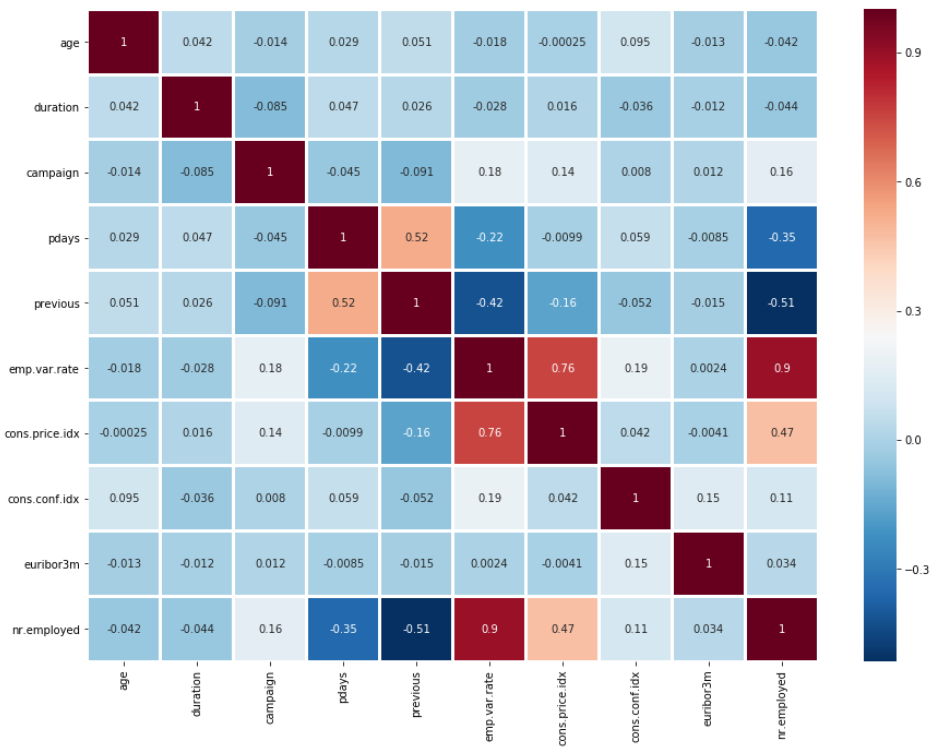
U setu je uočeno da postoje nedostajući podaci. Numerički atributi:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | duration | campaign | pdays | previous | emp.var.rate | cons.price.idx | cons.conf.idx | euribor3m | nr.employed |
| 0 | 3 | 5 | 0 | 2 | 3 | 2 | 2 | 2 | 1 |

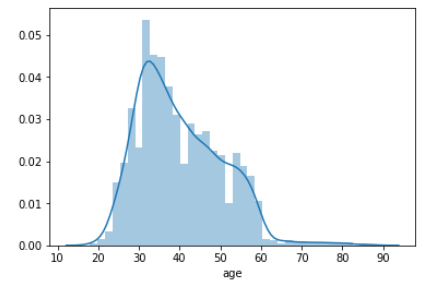
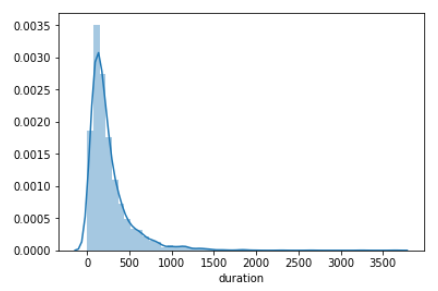
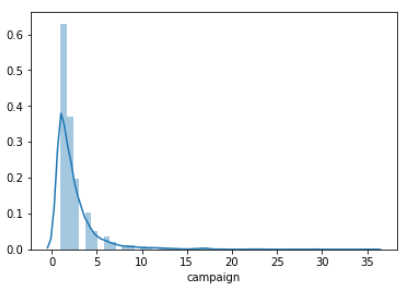
Kategorički atributi nemaju NaN vrednosti ali su uočene (kao što piše u inicijalnoj specifikaciji od autora data seta) da neki imaju „*unknown*“ vrednosti koje su prikazane u tabeli ispod:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| marital | education | default | housing | loan |
| 0.27% | 4.05% | 19.50% | 2.55% | 2.55% |

Odrađena je imputacija nedostajućih podataka. U prilogu je matrica korelacije numeričkih atributa:



U prilogu su prikazane distribucije nekih važnijih atributa:

# Priprema podataka za prediktivno modelovanje i rasponi vresnosti

Analiza ekstremnih vrednosti nije egzaktna. Pored tehnika za pronalaženje pomenutih ekstremnih vrednosti kao sto su z-score ili box-plot analiza, ne mozemo u startu oznaciti podatke kao ekstreme bez da logički ne proverimo raspone vrednosti. Kao primer, susreli smo se sa slučajem kolone 'age' gde nam algoritmi indikuju da postoje ekstremi. Međutim raspon godina ja 'razuman' tako da su tretirani kao retki ali validni slučajevi i korišćeni su u daljoj analizi.

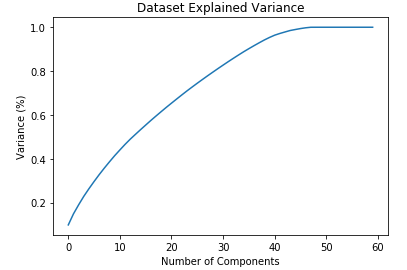
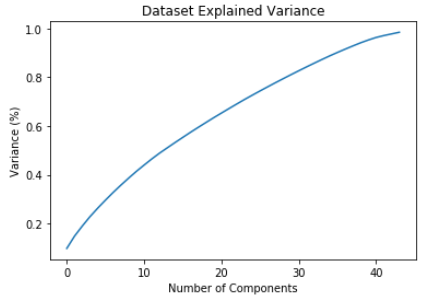
* Kolona '**age**' ima min 18 i max 88, i bez obzira što određene metode predstavljaju neke godine kao outlier-e, ne mozemo otpisati takve zapise. Iz razloga sto su min i max vrednosti razumne nećemo raditi uklanjanje outlier-a (ekstremnih vresnosti).
* Kolona '**duration**' ima minimalnu vrednost 0, a maksimalnu 3643 (u sekundama), što je oko 1h, i ovo bi trebalo dodatno proveriti ali u našoj zemlji je ograničenje dužine razgovora na max 1h što se poklapa sa mak trajanjem poziva u ovoj koloni. U pitanju je dužina poslednjeg razgovora.
* Kolona '**campaign**' ima minimalnu vrednosti 1 a maksimalnu vrednost 35. S obzirom da je ovo razuman broj poziva morali smo da uzmemo u obzir i te vrednosti.
* Kolona '**pdays**' ima minimalnu vrednost -1 a maksimalnu 21 što smatramo razumnim vrednostima. U pitanju je broj dana od poslednjeg kontakta. Vrednost -1 označava slučaj kada nije bilo kontakta.
* Kolona '**previous**' ima minimalnu vrednost -1 a maksimalnu 6, što smatramo razumnim vrednostima. U pitanju ukupan broj kontakata pre ove kampanje.
* Kolona '**emp.var.rate**' ima minimalnu vrednost -3.4 i maksimalnu 1.4, što smatramo razumnim vrednostima. U pitanju je stopa varijacije zaposlenosti.
* Kolona '**cons.price.idx**' ima minimalnu vrednost 92.201 i maksimalnu 94.767, što smatramo razumnim vrednostima. U pitanju je index potrošačkih cena.
* Kolona '**cons.conf.idx**' ima minimalnu vrednost -50.80 i maksimalnu 1.266 . Smatramo da ima jedan outlier koji ima z-score veci od 7. U pitanju je indeks poverenja potrošača.
* Kolona '**euribor.3m**' ima minimalnu vrednost 0.635 i maksimalnu vrednost 5099.100. Smatramo da ima jedan outlier čiji z-score je veći od 64. U pitanju je Euro Interbank Offered Rate.
* Kolona '**nr.employed**' ima minimalnu vrednost 4963.60 i maksimalnu vrednost 5228.10. Nakon box-plot analize i z-core filtera zaključeno je da nema ekstremnih vrednosti

U radu je korišćena strategija *dummy coding*[[3]](#footnote-3) za pretvaranje kategoričkih u numeričke podatke. Ovoj tehnici data je prednost u odnosu na tehniku *label encoder* [[4]](#footnote-4)kako bismo izbegli probleme sa regresionim algoritmima koji su osetljivi na sortirani brojeve (npr. 5 različitih 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 i nekim primerima je korišćena u ovom radu. Pre PCA transformacije podaci su skalirani.

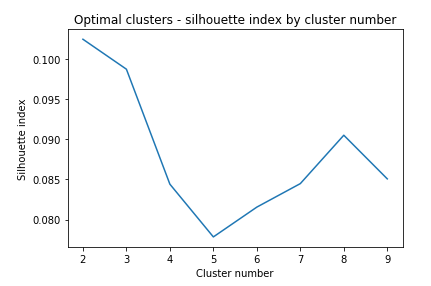
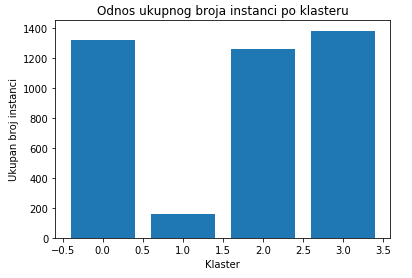
Sve komponente zajedno kumulativno nose 100% varijanse. Pre redukcije je isti broj komponenata kao i originalnih kolona (*slika 1*). Kao što se može primetiti, blizu 98% varijanse se može opisati sa oko 40 komponenti (*slika 2*).

Slika 1 – broj komponenti sa 100% kumulativne varijanse Slika 2 – broj komponenti sa oko 98% kumulativne varijanse

# Klasterovanje

U projektu za izvršenje klasterovanja korišćen je *K-Means* algoritam odnosno njegova unapređena verzija *K-means++,* koja ima poboljšane korake inicijalizacije. Takođe je korišćena mera evaluacije klastera *Silhouette Index Score* kako bismo odredili optimalan broj klastera (*slika 3*)*.*

Slika 3 – Silhouette index score odredjen za prvih 9 klastera Slika 4 – Ukupan broj instanci po klasteru (4 klastera)

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*). U prilogu su najznačajnije kolone centoida:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K | age | duration | campaign | pdays | previous | emp.var.rate | cons.price.idx | cons.conf.idx | euribor3m | nr.employed |
| 0 | 0.0139 | -0.0259 | 0.13478 | -0.1739 | -0.35018 | 0.70064 | 0.93676 | 0.386763 | 0.71457 | 0.525966 |
| 1 | 0.1999 | 0.21286 | -0.2896 | 4.3267 | 2.91949 | -1.3438 | -0.27872 | 0.470939 | -1.50208 | -1.91157 |
| 2 | -0.0042 | 0.02043 | -0.1970 | -0.1698 | 0.38528 | -1.27313 | -0.97243 | -0.5657 | -1.29858 | -1.12943 |
| 3 | -0.0323 | -0.0182 | 0.08411 | -0.1739 | -0.35108 | 0.646104 | 0.023754 | 0.09264 | 0.674138 | 0.746986 |

# Prediktivni modeli i njihova evaluacija

Za izdradu prediktivnih modela korišćeni su algoritmi mašinskog učenja Naïve Bayes, Logistička regresija, KNN, Decision three i Random Forest. Pre treniranja prediktivnih modela, podaci su normalizovani *MinMaxScaler()* funkcijom koja je rezultat sledeće jednačine , ili sa već pomenutim *StandardScore()* algoritmom. Kreirani su pajpovi (*eng. Pipeline*) i evaluirani su algoritmi pomoću metode cross\_validate[[5]](#footnote-5). U prilogu je tabela rezultata izvršavanja algoritama po sledećem redosledu: LogisticRegression, NaiveBayes, RandomForest, Knn, DecisionTree, RandomForest (balansirani podaci).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fit\_time | score\_time | test\_precision | train\_precision | test\_recall | train\_recall | test\_accuracy | train\_accuracy | test\_f1 | train\_f1 |
| 0.111272 | 0.006598 | 0.433547 | 0.450673 | 0.813118 | 0.86375 | 0.867109 | 0.874123 | 0.564533 | 0.592279 |
| 0.038092 | 0.006941 | 0.39758 | 0.39643 | 0.554516 | 0.551551 | 0.86155 | 0.863636 | 0.460456 | 0.461246 |
| 0.488958 | 0.023283 | 0.684345 | 1 | 0.285376 | 1 | 0.908737 | 1 | 0.397461 | 1 |
| 0.081918 | 0.158324 | 0.535952 | 0.827447 | 0.253118 | 0.450646 | 0.89729 | 0.931915 | 0.336688 | 0.58338 |
| 0.264135 | 0.012851 | 0.977644 | 1 | 0.881341 | 0.996486 | 0.933321 | 0.998243 | 0.888558 | 0.998239 |

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[[6]](#footnote-6)*, kako bismo dobili balansirane izlazne klase. Sledeći izazov je izabrati prave parametre algoritama i za tu svrhu korišćena je tehnika *GridSearch[[7]](#footnote-7).* Nakon toga tražene su kolone koje imaju najveći značaj (*eng. Feature importance*) pomoću metode *SelectKBest*[[8]](#footnote-8) *(slika 6)*.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **FeatureName** | duration | euribor3m | nr.employed | age | emp.var.rate | cons.conf.idx | campaign | cons.price.idx | pdays | poutcome\_success |
| **FeatureImportance** | 0.293 | 0.098 | 0.071 | 0.059 | 0.046 | 0.034 | 0.029 | 0.028 | 0.024 | 0.0167 |

# Zaključak

U prvom podskupu se dobijaju najbolje performanse a najlošije su u drugom jer taj klaster ima manje od 200 instanci pa je samim tim nije uspeo da generalizuje model i bude dobar na test podacima (pogledati tabelu u prošlom poglavlju). Performanse se poboljšavaju kada se algoritmi treniraju na kompletnom setu. Najveću prediktivnu moć imaju atributi *Duration, Euribor3m, nr.employed, age, emp.var.rate, cons.conf.idx, campaign, cons.price.idx, pdays* i *poutcome.cuccess*. Korišćenjem grid search metode pronađeni su optimalne vrednosti za određene parametre modela, koji dodatno poboljšavaju performanse. Tako optimizovani odnosno podešeni parametri daju bolje performanse od defaultnih. Potrebno je ispitati dodatno sve ostale parametre i testirati njihove performanse metodom grid search i pronaći najbolje konfiguracije za modele za dati set podataka. Postoji verovatnoća da je model Random Forest koji je treniran sa jako malim podacima pretreniran.

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. 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-3)
4. Label encoder – tehnika koja konvertuje kategoričke podake u sortirane brojeve. [↑](#footnote-ref-4)
5. 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-5)
6. Tehnika kreiranja novih zapisa u cilju balansiranja izlazne klase. [↑](#footnote-ref-6)
7. Tehnika pretrage optimalni parametara koja radi po metodi probaj sve i vrati najbolji. [↑](#footnote-ref-7)
8. Metoda koja selektuje atribute koji imaju najbolji skor. [↑](#footnote-ref-8)