Klasifikacija u bankovnom marketingu

Seminarski rad u okviru kursa
Istraživanje podataka 1
Matematički fakultet

Aleksandra Nikšić
mi16072@matf.bg.ac.rs

9. maj 2019.

Sažetak

Sadržaj

1 Uvod

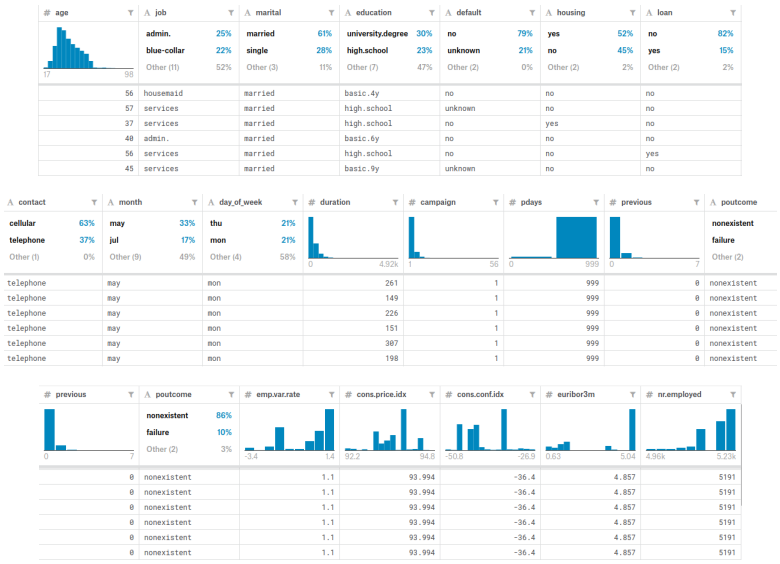
Nadgledano mašinsko učenje karakteriše se time da su za sve podatke poznate vrednosti ciljne promenljive. Problemi ove vrste mašinskog učenja se uglavnom mogu razvstati u jednu od dve grupe - probleme regresije i probleme klasifikacije. U nastavku ćemo se baviti problemom klasifikacije. Problem klasfikacije je problem razvrstavanja nepoznate instance u jednu od unapred ponuđenih kategrija - klasa. Svaka instanca se može predstaviti skupom atributa. Cilj je odredjivanje vrednosti atributa klase na osnovu preostalih atributa instance.
Postoji veliki broj algoritama kojima se ovaj problem rešava. Neki od njih su:

- metoda potpornih vektora (SVM)
- C&RTree(Classification And Regression Trees)
- stabla odlučivanja
- C5.0
- k najbližih suseda (KNN)

Softverski paket **SPSS Modeler** korišćen je za primenu prva 4, a **Python** za primenu KNN algoritma na dati skup podataka.

1.1 Skup podataka

Korišćeni skup podataka dostupan je na Kaggle sajtu: https://www.kaggle.com/henriqueyamahata/bank-marketing pod imenom Bank Marketing.
Na osnovu tog skupa vršićemo pomenutu predikciju.

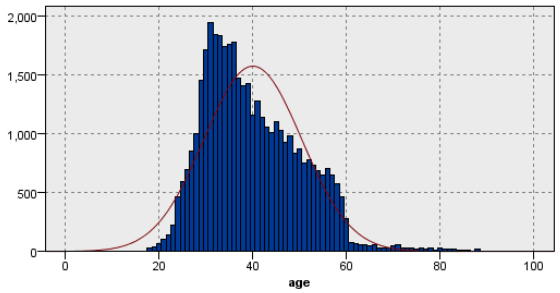


Slika 1: Uvid u tabelu

1.2 Analiza skupa

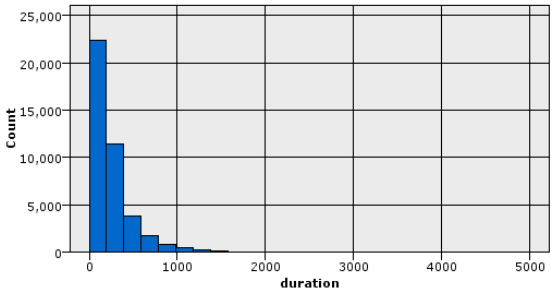
Skup se sastoji od 21 atributa, postoji 41188 slogova, pri čemu nema nedostajućih vrednosti. Skup podataka sadrži 10 numeričkih atributa, 11 kategoričkih. Data set je čist i smatram da je već bio preprocesiran.
Upoznaćemo se sa pojmom koji je centar našeg istraživanja:
Oročni depozit - deponovana (uložena) novčana sredstva kod banke ili drugog lica sa unapred ugovorenim rokom oročenja i fiksnom kamatnom stopom. Oročeni depozit se ne može razročiti sem ukoliko su saglasne obe strane. Depozit može biti oročen kratkoročno (do 12 meseci) i dugoročno (preko 12 meseci).

Na slici je prikazan histogram broja ispitanika u zavisnosti od uzrasta. Najveći broj ispitanika ima 31 godinu.



Slika 2: Uzrast ispitanika

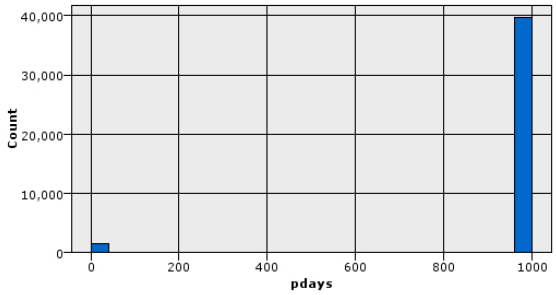
Sledeća slika daje uvid u trajanje poziva prilikom anketiranja. Ovo je jedan od najbitnijih atributa. Zanimljivost: Najveći je udeo poziva koji traju 0 sekundi. Ukoliko je trajanje poziva 0, sigurni smo da je odgovor NE.



Slika 3: Trajanje poziva

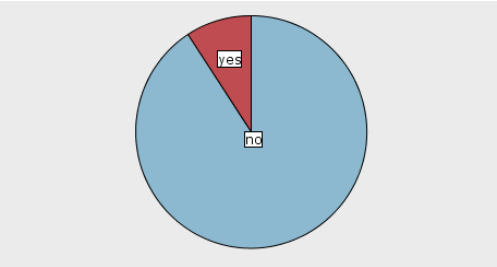
Obratimo pažnju na atribut *pdays*, koji predstavlja broj dana od poslednjeg kontaktiranja potencijalnog klijenta. On uglavnom ima vrednost 999, što znači da je osoba kontaktirana prvi put. Na prvi pogled, možemo ga izostaviti iz analize jer je skoro sigurno svaki klijent prvi put pozvan.

Ovaj atribut međutim ne možemo isključiti jer imamo češću pojavu pri-

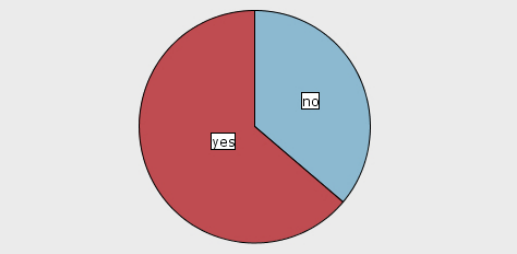


Slika 4: Histogram atributa *pdays*

stanka na ponudu ako je osoba pozvana 2. ili 3. put što nas dovodi do zaključka da vredi iznova zvati već kontaktirane osobe. Na slikama 5 i 6 možemo videti i vizuelni prikaz iz kog je jednstavno zaključiti da se broj pristanaka povećava ukoliko je ostvareno višestruko kontaktiranje osobe ($pdays < 999$).



Slika 5: Odgovori ispitanika za vrednost $pdays = 999$



Slika 6: Odgovori ispitanika za vrednost $pdays < 999$

2 Primena algoritama klasifikacije

Isprobano je 5 algoritma, među kojima su pojedini ispitani sa, a pojedini bez upotrebe glavih komponenti. Takođe, algoritam koji koristi

stabla odlučivanja pokrenut je i sa i bez najuticajnijeg prediktora.
Na slici 7 prikazano je koliko komponenti je dovoljno za određeni procenat pokrivenosti skupa. Sa samo 5 novih komponenti moguće je pokriti 83.122% skupa, a već sa 6 pokrivenost je čak 91.540%.

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.896	38.958	38.958	3.896	38.958	38.958
2	1.359	13.589	52.547	1.359	13.589	52.547
3	1.078	10.777	63.324	1.078	10.777	63.324
4	1.050	10.504	73.827	1.050	10.504	73.827
5	.929	9.294	83.122	.929	9.294	83.122
6	.842	8.418	91.540	.842	8.418	91.540
7	.425	4.247	95.787			
8	.386	3.857	99.644			
9	.025	.248	99.893			
10	.011	.107	100.000			

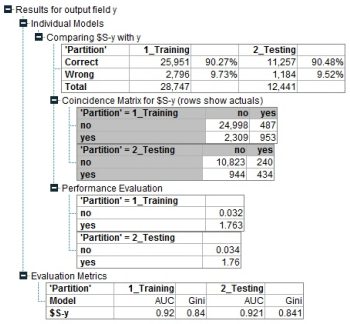
Slika 7: PCA faktori

2.1 SVM

Prvi algoritam koji ćemo primenti je SVM (support vector machines). On je zasnovan na vektorima i pronalaženju hiperravni koja će razdvojiti elemente koji pripadaju jednoj klasi od onih koji pripadaju drugoj.
Algoritam ćemo izvršiti sa korišćenjem analize glavnih komponenti i bez nje, a potom uporediti rezultate.

2.1.1 SVM sa PCA

Koristimo se analizom glavnih komponenti, tj. novih atributa, od kojih je svaki linearna kombinacija originalnog skupa. Prednost ove analize leži u tome što radimo sa manjim skupom atributa.



Dobijeni rezultati govore da algoritam uspešno klasifikuje u 90.27% slučajeva nad trening, a 90.48% nad test podacima. Interesantno je da imamo praktično istu preciznost nad oba skupa.

2.1.2 SVM bez PCA

Medjutim, i pored povećanja efikasnosti smanjenjem broja atributa, javlja se jedna zanimljivost. Algoritam SVM bez korišćenja PCA daje procenat uspešnosti 94.85% nad trening skupom, dok nad test skupom daje nešto manju vrednost nego sa PCA, 90.21%.

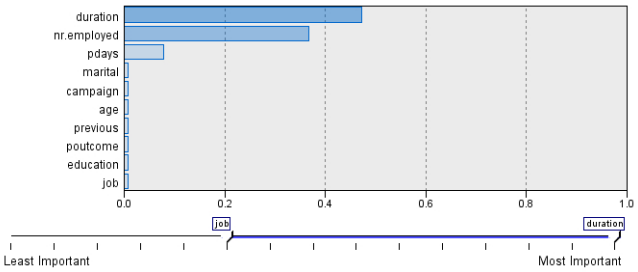
Results for output field y

Comparing $S-y with y

'Partition'	1_Training		2_Testing	
Correct	27,266	94.85%	11,223	90.21%
Wrong	1,481	5.15%	1,218	9.79%
Total	28,747		12,441	

2.2 C&RTree

Algoritam C&RTree koristi se stablima pri rešavanju problema koji je ujedno klasifikacione i regresione prirode.
Na slici 8 dat je prikaz značajnosti atributa za rad algoritma. Atributi



Slika 8: Važnost prediktora u algoritmu C&RTree

koji najviše utiču na efekat ovog alogritam su: *duration*, *nr.employed* i *pdays*.
Preciznost na trening skupu: 91.41.%
Preciznost na test skupu: 90.99%.
Na slici 9 možemo videti matricu konfuzije ovog algoritma:

'Partition' = 1_Training	no	yes
no	24,596	889
yes	1,581	1,681
'Partition' = 2_Testing	no	yes
no	10,650	413
yes	708	670

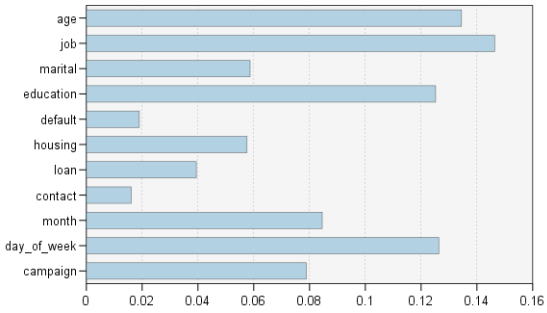
Slika 9: Matrica konfuzije C&RTree algoritma

2.3 RandomTree

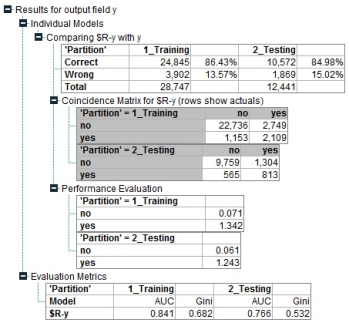
Krećući se od korena drveta odlučivanja, slog se podlaže test uslovu nakon čega ulazimo u granu koja odgovara dobijenom rezultatu. Ukoliko smo na tom putu naišli na unutrašnji čvor, ponovićemo prethodni postupak, dok nailaskom na list slogu dodeljujemo klasu pridruženu tom listu. U primeni ovog algoritma, razlikovaćemo dva slučaja. Naime, primećeno je da najveći uticaj od svih atributa ima atribut *duration*, trajanje poziva. Ispitaćemo uspešnost klasifikacije sa i bez njega.

2.3.1 Stablo odlučivanja, bez najuticajnijeg atributa

Na slikama 10 i 11, redom, videćemo važnost atributa i rezultate algoritma.



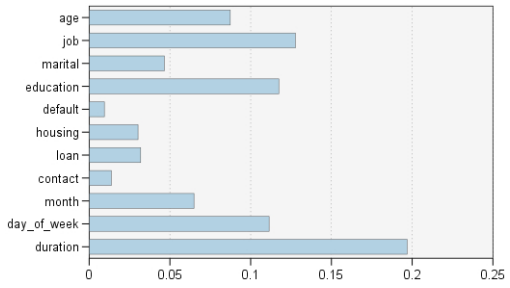
Slika 10: Uticaj atributa bez *duration* (Predictor importance)



Slika 11: Rezultati algoritma

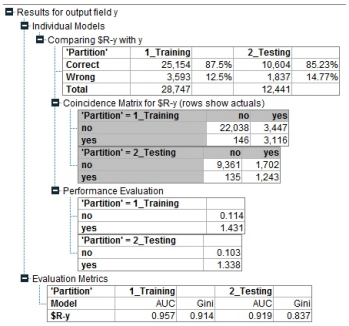
2.3.2 Stablo odlučivanja, sa najuticajnijim atributom

Na slikama 12 i 13, redom, videćemo važnost atributa i rezultate algoritma.



Slika 12: Uticaj atributa sa *duration* (Predictor importance)

Primetićemo da je uspešnost algoritma manja za približno procenat ako ne koristimo atribut *duration*.



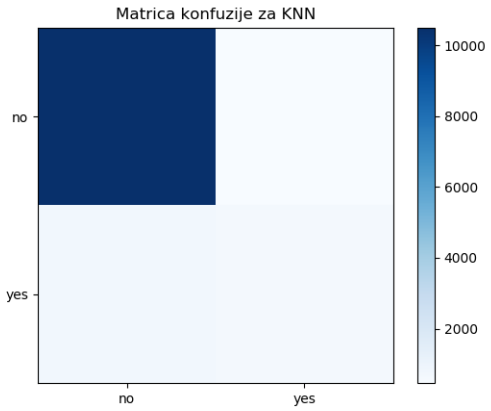
Slika 13: Rezultati algoritma

2.4 C5.0 sa PCA

Ovaj algoritam koristi entropiju kao meru i odnos informacione dobiti kao kriterijum podele. On formira n-arno stablo, a kada stignemo do lista za klasu u njemu biramo najbrojniju od svih klasa. Algoritam je poput stabla odlučivanja i SVM algoritma podvrgnut prethodnoj PC analizi pre same primene.
Preciznost na trening skupu: 90.44%.
Preciznost na test skupu: 90.31%.

2.5 KNN

U algoritmu k najbližih suseda, element se klasifikuje glasovima većine svojih suseda, tako da biva raspoređen u klasu koja je najčešća među njegovih k suseda. Broj komšija koji je izabran za testiranje je 5.
Preciznost algoritma nad trening podacima iznosi 90.51%, dok je nad test skupom 93.03%.



Slika 14: Matrica konfuzije KNN algoritma

3 Zaključak i diskusija rezultata

Rezultati svih algoritama su veoma dobri. Kao najbolji na **trening** setu pokazao se SVM bez PCA (94.85%), dok je najlošije rezultate dalo drvo odlučivanja bez najuticajnijeg atributa (86.43%). Najveći procenat uspešnosti klasifikacije na **test** skupu imao je KNN (93.03%), a najmanji drvo odlučivanja (84.98%). Ako posmatramo prosečnu uspešnost na oba skupa istovremeno, možemo reći da je najbolji od korišćenih upravo SVM bez PCA (92.53%), a najgori drvo odlučivanja bez najuticajnijeg atributa (85.705%).

Koristeći PCA u slučaju SVM algoritma bezmalo smo dobili podudaranje preciznosti na skupovima za trening i test.

Kod algoritama SVM sa PCA i KNN uočeno je povećanje procenta uspešnosti nad test skupom u odnosu na trenirane podatke.

Poredeći rad algoritma sa i bez uticajnog atributa, dobili smo razliku u procentu efikasnosti.

Posmatrajući vrednosti atributa *pdays* i odgovora ispitanika utvrdili smo da će kontaktiranje iste osobe više puta uroditi plodom i povećati verovatnoću pristanka na oročeni depozit.