Seminarski rad u okviru kursa Istraživanje podataka 1
Matematički fakultet, Univerziteta u Beogradu

Analiza skupa Bank Customer Survey – Marketing for Term Deposit metodom klasifikacije

Tamara Ivanović 462/2018

mi14262@alas.matf.bg.ac.rs

13.08.2019. Beograd

1 Uvod

U ovom seminarskom radu biće prikazan proces klasifikacije nad podacima preuzetih sa veb sajta Kaggle (https://www.kaggle.com/sharanmk/bank-marketing-term-deposit). Skup podataka predstavlja informacije o klijentima jedne banke i oročenom depozitu prikupljenih prilikom telefonske ankete (marketinške kampanje). Cilj kampanje je da se privuku klijenti da uzmu oročeni depozit, samim tim je ciljni atribut istraživanja procena da li je klijent odabrao oročeni depozit.

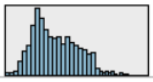
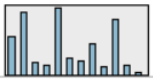
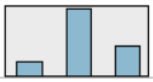
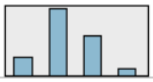
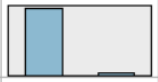
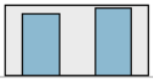
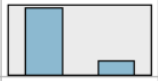
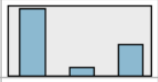
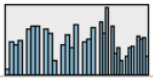
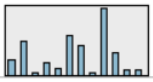
2 Analiza i pretprocesiranje

Podaci se nalaze u tabeli koja sadrži 45211 instanci. Svaka instanca predstavlja informacije o jednom klijentu i opisana je pomoću 17 atributa. Lista atributa:

- **age** – godine klijenta
- **job** –zanimanje klijenta, kategoički atribut (12 kategorija, uključujući unknown)
- **marital** – bračno stanje, kategorički atribut (married, single, divorced)
- **education** – nivo obrazovanja, kategorički atribut(primary, secondary, tertiary, unknown)
- **default** – da li klijent ima neotplaćeni kredit (da li ima dug), binarni atribut (yes, no)
- **balance** – prosečna godišnja zarada u evrima
- **housing** – da li klijent ima stambeni kredit, binarni atribut (yes, no)
- **loan** – da li klijent ima lični zajam, binarni atribut (yes, no)
- **contact** – način na koji je klijent kontaktiran, kategorički atribut (unknown, telephone, cellular)
- **day** – poslednji dan kada je klijent kontaktiran, numerički atribut
- **month** – mesec u kom je klijent poslednji put kontaktiran, katetorički atribut (jan, feb, ... , nov, dec)
- **duration** – dužina tog razgovora u sekundama, numerički atribut
- **campaign** - koliko puta je ovaj klijent kontaktiran tokom ove kampanje, uključujući poslednji razgovor
- **pdays** – broj dana koji je protekao između prethodne i ove kampanje za datog klijenta (-1 ukoliko je klijent prvi put kontaktiran)
- **previuous** - broj poziva upućenih klijentu pre ove kampanje
- **poutcome** – ishod prethodne marketinške kampanje, kategorički atribut (unknown, other, failure, success)
- **y** – da li se klijent prijavio za oročeni depozit, binarni atribut (0, 1)

Pre samog pretprocesiranja podataka potrebno je proveriti šta treba da se uradi sa podacima. Prvi deo je urađen u programskom jeziku Python gde je provereno postojanje null vrednosti i uočeno da ih nema

u ovom skupu. Analizom korelacije kategoričkih atributa sa ciljnim atributom uočava se da pojedini atributi imaju vrednosti *unknown* koje je potrebno kategorisati, a ostale statističke analize nam pomažu koje atribute možemo da uklonimo iz razmatranja. U SPSS modeleru je urađeno dodatno pretprocesiranje prethodno smanjenog skupa.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
age		Continuous	18	95	40.936	10.619	0.685	--	45211
job		Nominal	--	--	--	--	--	12	45211
marital		Nominal	--	--	--	--	--	3	45211
education		Nominal	--	--	--	--	--	4	45211
default		Flag	--	--	--	--	--	2	45211
balance		Continuous	-8019	102127	1362.272	3044.766	8.360	--	45211
housing		Flag	--	--	--	--	--	2	45211
loan		Flag	--	--	--	--	--	2	45211
contact		Nominal	--	--	--	--	--	3	45211
day		Continuous	1	31	15.806	8.322	0.093	--	45211
month		Nominal	--	--	--	--	--	12	45211

Slika 1 Prikaz grafika i statistika atributa

2.1 Pretprocesiranje u programskom jeziku Python

Kao što je rečeno, pojedini kategorički atributi imaju instance sa vrednostima *unkown*. U atributu *job* korelacija ‘unknown’ je 0.118056, a ‘self’ je 0.118429 i samim tim su instance prebačene u kategoriju 'self'. Slično, kod atributa *education* korelacija 'unknown' je 0.135703 i 'tertiary' je 0.150064 to je dovoljno blizu da se kategoriše kao'tertiary'. Atribut *poutcome* ima preveliki broj unknown vrednosti i samim tim ne može se lako odrediti kojoj bi se kategoriji te vrednosti dodale, a i sam atribut nam nije od velikog značaja. Samim tim taj atribut će biti uklonjen. Atribut *contact* ima mali broj unknown vrednosti, ali nam način komunikacije nije od velikog značaja i nećemo ga dalje posmatrati.

job	y
student	0.286780
retired	0.227915
unemployed	0.155027
management	0.137556
admin	0.122027
self	0.118429
unknown	0.118056
technician	0.110570
services	0.088830
housemaid	0.087903
entrepreneur	0.082717
blue	0.072750

education	y
tertiary	0.150064
unknown	0.135703
secondary	0.105594
primary	0.086265

index	y
y	1.000000
duration	0.394521
pdays	0.103621
previous	0.093236
campaign	0.073172
balance	0.052838
day	0.028348
age	0.025155

Slika 2 Korelacije atributa a) job i y; b) education i y; c) numeričkih atributa

Posmatranjem koeficijenta korelacije uočava se da atribut duration ima najveći koeficijent i da je samim tim on najznačajniji predictor. Analizom prosečnih vrednosti atribtuta koji imaju numeričke vrednosti može se zaključiti da *day* ima sličnu srednju vrednost za oba ishoda i time nam je taj atribut nebitan.

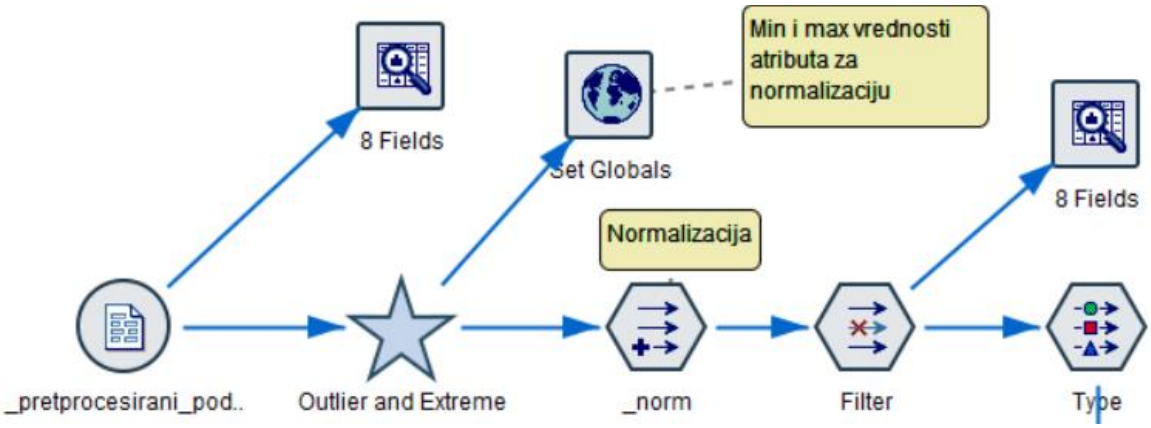
y	age	balance	day	duration	campaign	pdays	previous
0	40.838986	1303.714969	15.892290	221.182806	2.846350	36.421372	0.502154
1	41.670070	1804.267915	15.158253	537.294574	2.141047	68.702968	1.170354

Slika 3 Prosečne vrednosti atributa

Atributi *duration, pdays, previous, campaign, balance, job* i *education* su prosleđeni u novi csv fajl koji će dalje biti korišćen u SPSS, a za dalji rad u Python-u ostavljamo samo numeričke podatke i izbacujemo job i education. Nad tim podacima izvršena je normalizacija pomoću funkcije *MinMaxScaler()* i pomoću *train_test_split()* su podaci podeljeni na trening i test skup u odnosu 75:25.

2.2 Pretprocesiranje u SPSS modeleru

Kao ulazni fajl odabran je *pretprocesirani_podaci.csv* koji sadrži attribute koji su pročišćeni kroz programski jezik Python. Sastoji se od 8 polja, gde nijedno ne sadrži unknown vrednosti. Za početak pomoću čvora Data Audit izvršen je pregled atributa koji postoje i pravljenjem Outlier and Extreme super čvora pročistili podatke od autlajera i ekstremnih vrednosti. Kod svih atributa koji su imali autlajere te vrednosti su postavljene na granicu, a jedino kod atributa *previous* su ektremne vrednosti odbačene. Nakon ovoga izračunate su minimalne i maksimalne vrednosti atributa i te vrednosti su sačuvane kao globalne kako bi se podaci normalizovali. Svi atributi sa numeričkim vrednostima su primenom formule u čvoru *Derive* normalizovani. S obzirom da su normalizovane vrednosti sačuvane kao nove kolone primenom *Filter* čvora su izbačeni atributi sa starim vrednostima i u skupu su ostale samo normalizovane. Poslednji korak u pripremi podataka je podela skupa na trening i test skup. Ovo je urađeno korišćenjem čvora *Partition,* gde je odabrano da 70% čini trening i 30% test skup.



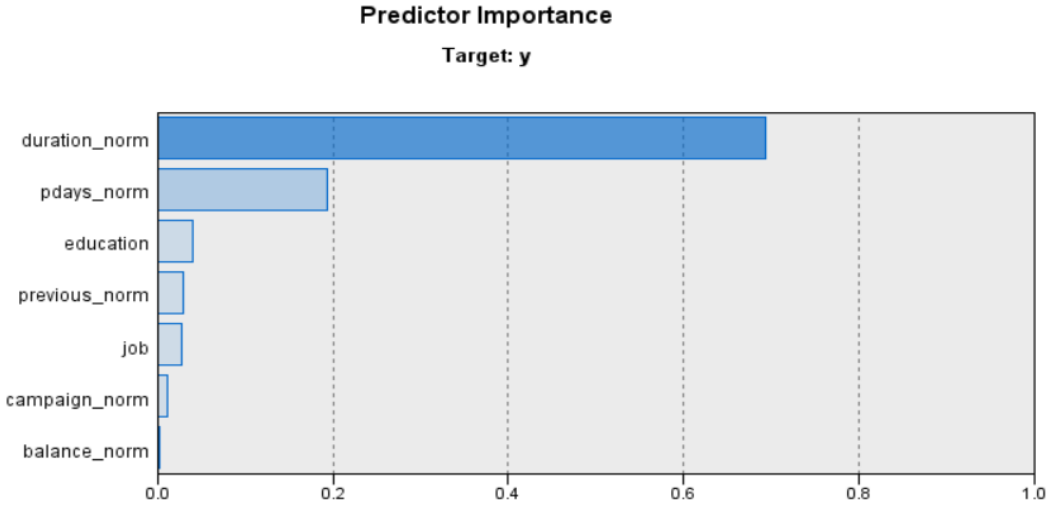
Slika 4 Shema pretprocesiranja u SPSS modeleru

3 Klasifikacija

Cilj ovog rada je da primenom algoritama klasifikacije i analizom dobijenih rezultata dođemo do što boljeg modela za klasifikaciju da li će klijent uzeti oročeni deposit ili ne. Kako je ovo problem koji zbog svog ishoda spada u binarnu klasifikaciju primenjeni su upravo algoritmi koji dobro rade podelu na 2 skupa. U ovom poglavlju će biti objašnjen svaki od algoritama koji je korišćen zajedno sa analizom dobijenih rezultata.

3.1 C5.0

Prilikom poziva čvora za algoritam C5.0 odabrano je da se generiše drvo odlučivanja. Dobijeno je drvo dubine 13, posmatrano je 31455 instanci i preciznost analize je 90.443%.

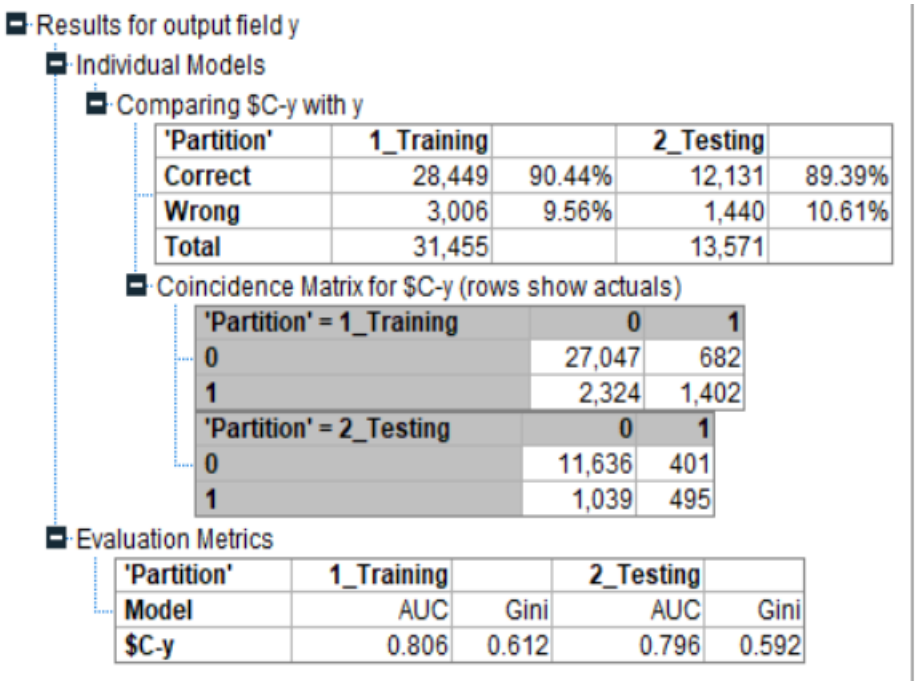


Slika 5 Važnost atributa, algoritam C5.0

Nodes	Importance
balance_norm	0.0024
campaign_norm	0.0121
job	0.0274
previous_norm	0.0295
education	0.0402
pdays_norm	0.1941
duration_norm	0.6943

Tabela 1 Važnost atributa u algoritmu C5.0

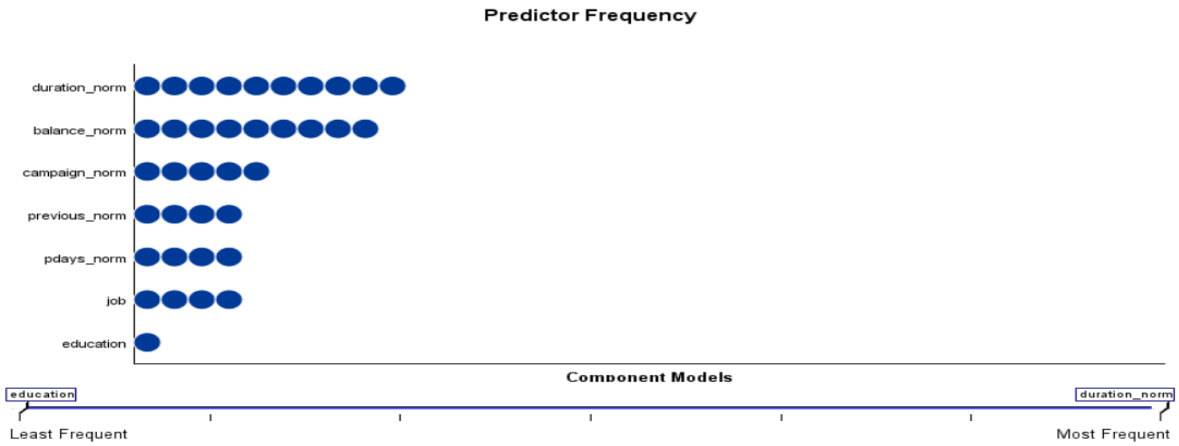
Na slici 5 se vidi da je najbitniji atribut *duration*, dok su prihodi i koliko je korisnik puta kontaktiran nebitni podaci. Čvor *Analysis* nam daje detaljniju analizu ovog modela. Trening skup daje 90.44% tačnosti, a test skup 89.39% tačnosti. S obzirom da je razlika veoma mala imamo model koji nije preprilagođen. Posmatranjem AUC indeksa koji je 0.806 za trening i 0.796 za test skup dobijamo da je model dobar jer je razlika mala, a dovoljno je blizu 1 (što čini idealan model).



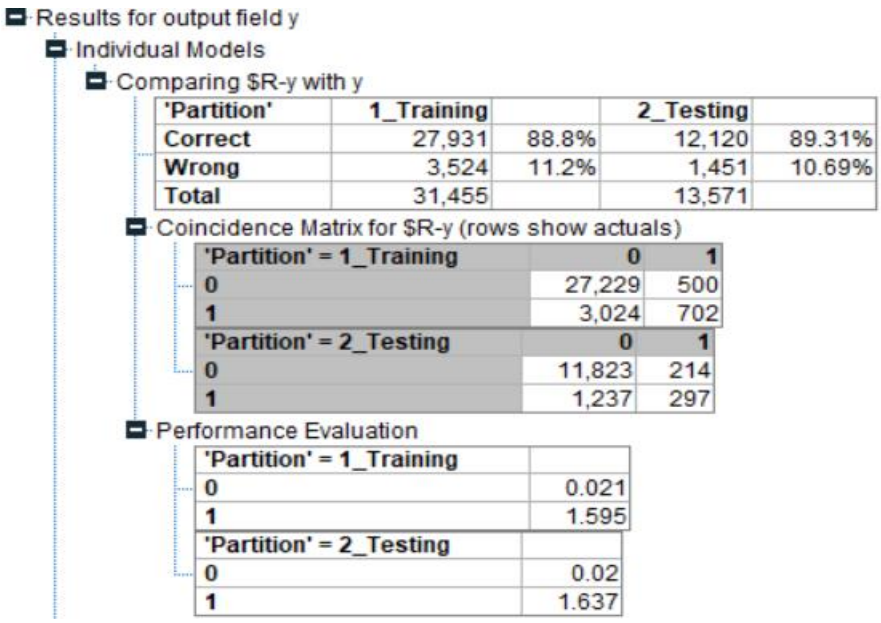
Slika 6 Matrica konfuzije algoritam C5.0

3.2 CART

Pokretanjem algoritma sa minimalnim uslovima – maksimalna dubina stable 4 i korišćenje Gini indeksa dobija je model koji koristi samo 2 prediktora. Radi nalaženja modela koji će imati bolju stabilnost izabrano je *Enhance model stability.* Svakako se dobija da je najbitniji predictor *duration,* međutim za razliku od prethodnog algoritma za njim je *balance.* Učestalost prediktora u ovom modelu se vidi na slici 7. Takođe vidimo da bez obzira koliko predoktora uzeli u razmatranje i koliko čvorova imali preciznost modela je uvek približno 88%. I u ovom algoritmu je preciznost trening i test skupa veoma blizu i samim tim ni ovo nije preprilagođen model.



Slika 7 Učestalost prediktora, algoritam CART

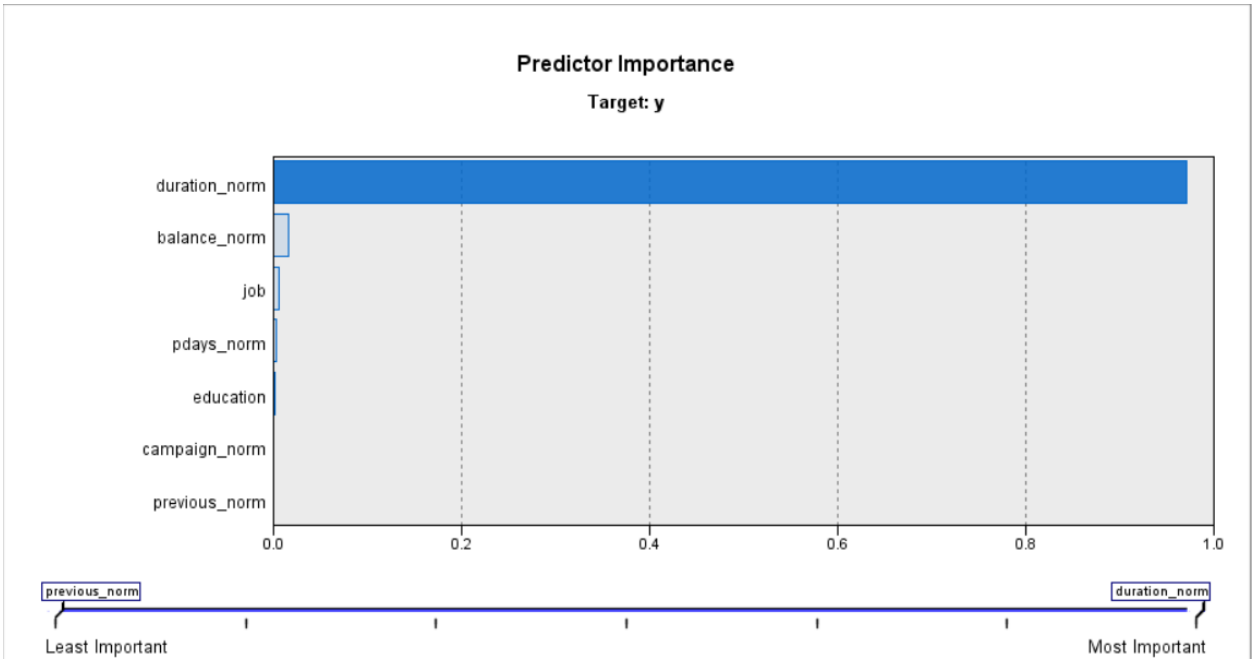


'Partition'	1_Training		2_Testing	
Correct	27,931	88.8%	12,120	89.31%
Wrong	3,524	11.2%	1,451	10.69%
Total	31,455		13,571	

'Partition' = 1_Training	0	1
0	27,229	500
1	3,024	702
'Partition' = 2_Testing	**0**	**1**
0	11,823	214
1	1,237	297

'Partition' = 1_Training	
0	0.021
1	1.595
'Partition' = 2_Testing	
0	0.02
1	1.637

Slika 8 Matrica konfuzije, algoritam CART



Slika 9 Važnost prediktora, algoritam CART

Component Model Details

Model	Accuracy	Method	Predictors	Model Size (Nodes)	Records
1	88.8%	C&RT	6	11	31,455
2	88.7%	C&RT	2	5	31,455
3	88.6%	C&RT	2	5	31,455
4	88.8%	C&RT	7	5	31,455
5	88.6%	C&RT	2	5	31,455
6	88.8%	C&RT	3	7	31,455
7	88.7%	C&RT	1	5	31,455
8	88.8%	C&RT	6	9	31,455
9	88.8%	C&RT	2	5	31,455
10	88.8%	C&RT	6	7	31,455

Slika 10 Detalji modela CART algoritma u odnosu na broj prediktora

3.3 Algoritam k najbližih suseda (KNN)

Kod ovog algoritma smo gledali preciznost modela. U tom slučaju se automatski bira najbolji broj suseda iz većeg skupa, dok za računanje udaljenosti koristi važnost prediktora. k je između 3 i 8, a za računanje udaljenosti je korišćeno Menhetn rastojanje (rastojanje gradskih blokova). Svi prediktori su jednako važni.

Posmatrajući nivo greške dobija se da je za k=6 i k=8 najmanja greška. S obzirom da je ovde bio cilj da pronađemo model koji je što precizniji rezultati od 90.09% za trening skup i 90.16% za test skup su odlični i daju nam model koji dobro klasifikuje, a čekanje od manje od 2 minuta, za obradu 31455 instanci za ovakvu preciznost nije dugo.

Slika 11 Greška u zavisnosti od broja suseda, algoritam KNN

Results for output field y

Individual Models

Comparing $KNN-y with y

'Partition'	1_Training		2_Testing	
Correct	28,339	90.09%	12,235	90.16%
Wrong	3,116	9.91%	1,336	9.84%
Total	31,455		13,571	

Coincidence Matrix for $KNN-y (rows show actuals)

Confidence Values Report for $KNNP-y

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$KNN-y	0.916	0.831	0.917	0.834

Slika 12 Matrica konfuzije, algoritam KNN

3.4 QUEST

Ovaj algoritam nam generiše binarno stablo. Za date podatke koliku god da stavimo maksimalnu dubinu stabla dobijamo stablo dubine 2. Kao najbitniji atribut je obabran *duration* dok su svi ostali atributi jednako bitni, ali znatno manje u odnosu na duration. Sama preciznost algoritma je solidna, slična za trening i za test skup. Međutiim AUC je 0.637, a kako je ova vrednost bliža 0.5 može se reći da ovaj model naginje ka lošem klasifikatoru.

Results for output field y

Individual Models

Comparing $R-y with y

'Partition'	1_Training		2_Testing	
Correct	27,929	88.79%	12,123	89.33%
Wrong	3,526	11.21%	1,448	10.67%
Total	31,455		13,571	

Coincidence Matrix for $R-y (rows show actuals)

Performance Evaluation

Confidence Values Report for $RC-y

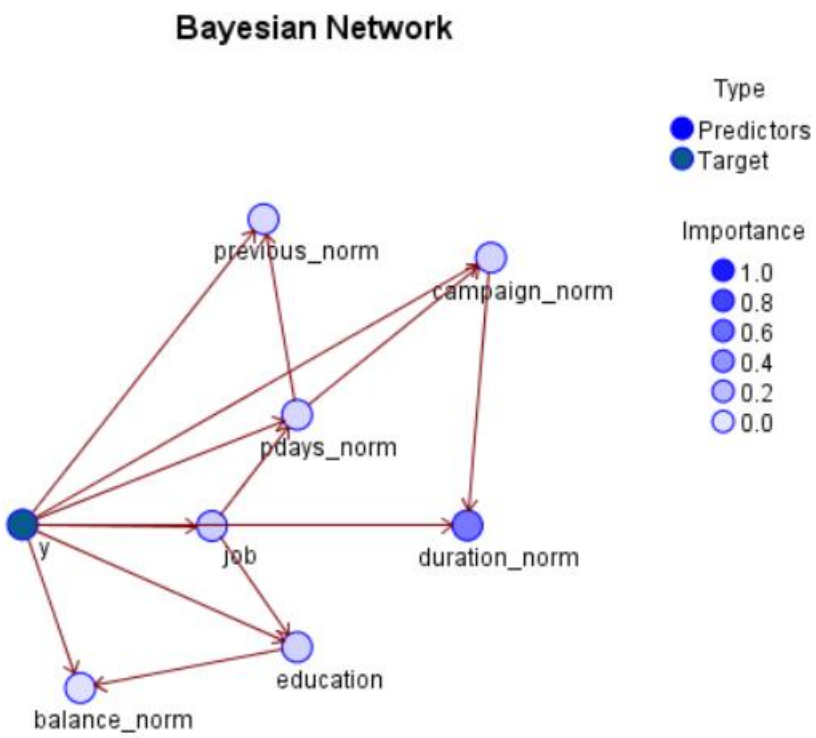
Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$R-y	0.637	0.275	0.645	0.289

Slika 13 Matrica konfuzije, algoritam QUEST

3.5 Bajesove mreže (Bayes net)

Ovaj model koristi Bajesovu statistiku. Posmatra *duration* a zatim *job*. Za posao izračunava uslovne verovatnoće za svaku instancu. Takođe spade u dobar model na osnovu matrice konfuzije.



Slika 14 Bajesova mreža

- Results for output field y
  - Individual Models
    - Comparing $B-y with y

'Partition'	1_Training		2_Testing	
Correct	27,981	88.96%	12,106	89.2%
Wrong	3,474	11.04%	1,465	10.8%
Total	31,455		13,571	

    - Coincidence Matrix for $B-y (rows show actuals)
    - Performance Evaluation
    - Confidence Values Report for $BP-y
  - Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$B-y	0.859	0.718	0.852	0.703

Slika 15 Matrica konfuzije, algoritam Bajesova mreža

3.6 Naivan Bajesov algoritam

Ovaj algoritam je primenjen u Python-u pomoću biblioteke *sklearn.naive_bayes*, a korišćena je Gausova formula tj. za verovatnoću je korišćena verovatnoća normalne raspodele.

Kako bi mogao da se prikaže izveštaj klasifikacije potrebno je da ciljni atribut bude kategorički, zato je prvo promenjeno da je 1 yes, a 0 no. Ovaj algoritam je veoma brz za izvršavanje i daje dobru preciznost. Dobijena je preciznost od 0.87.

```
Naivan Bajesov algoritam
Matrica konfuzije
[[9437  544]
 [ 898  424]]
Preciznost 0.8724232504644784


Izvestaj klasifikacije
              precision    recall  f1-score   support

         yes       0.91      0.95      0.93      9981
          no       0.44      0.32      0.37      1322

    accuracy                           0.87     11303
   macro avg       0.68      0.63      0.65     11303
weighted avg       0.86      0.87      0.86     11303
```

Slika 16 Matrica konfuzije i izveštaj klasifikacije, naivan Bajesov algoritam

4 Zaključak

Svi algoritmi koji su primenjeni daju dobre rezultate, ostalo je uporetiti ih. Na osnovu matrica konfuzije koje su dobijene prilikom analize dobijaju se sledeće preciznosti:

Algoritam	Trening skup	Test skup
C5.0	90.44	89.39
CART	88.80	89.31
KNN	90.09	90.16
QUEST	88.79	89.33
Bajesova mreža	88.96	89.20
Naivni Bajes	-	87.24

Posmatranjem odnosa preciznosti za test skup dobija se da najbolje rezultate daje KNN algoritam.