

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

Tamara Ivanović, 462/2018

Seminarski rad u okviru kursa
Istraživanje podataka 1
Matematički fakultet

Avgust, 2019.

Uvod

- Klasifikacija skupa bank marketing term deposit
- Analiza i pretprocesiranje
- Klasifikacija u SPSS modeleru
- Klasifikacija u Pythonu
- Analiza dobijenih modela

Lista atributa

- **age** – godine klijenta
- **job** – zanimanje klijenta, kategorički 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, kategorički 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)
- **previous** – broj poziva upućenih klijentu pre ove kampanje
- **outcome** – ishod prethodne marketinške kampanje, kategorički atribut (unknown, other, failure, success)
- **y** – da li se klijent prijavio za oročeni depozit, binarni atribut (0, 1)

Slika 1: Atributi skupa

Pretprocesiranje u Python-u

- Nema *null* vrednosti
- *unknown* zamenjeno kod *education* i *job*
- Uklonjeni atributi *poutcome*, *contact* i *day*
- `MinMaxScaler()`, `train_test_split()`

	age	balance	day	duration	campaign	pdays	previous
y							
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 2: Prosečne vrednosti atributa

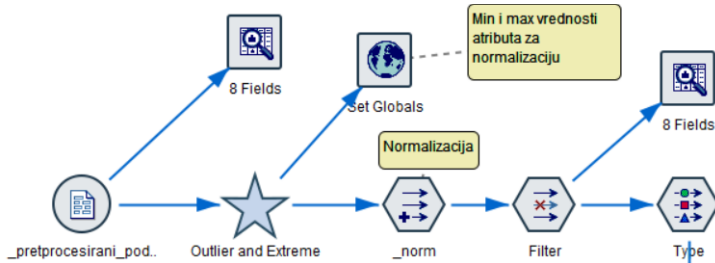
- *duration*, *pdays*, *previous*, *campaign*, *balance*, *job*, *education*

Pretprocesiranje u Python-u

	y				
job				index	y
student	0.286780			7	y
retired	0.227915			3	duration
unemployed	0.155027			5	pdays
management	0.137556			6	previous
admin	0.122027			4	campaign
self	0.118429			1	balance
unknown	0.118056			2	day
technician	0.110570			0	age
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		

Slika 3: Korelacije atributa

Pretprocesiranje u SPSS modeleru



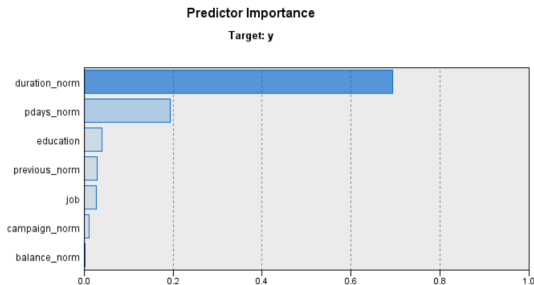
Slika 4: Shema pretprocesiranja u SPSS modeleru

Klasifikacija

- C5.0
- CART
- KNN
- QUEST
- Bajesove mreže
- Naivan Bajesov algoritam

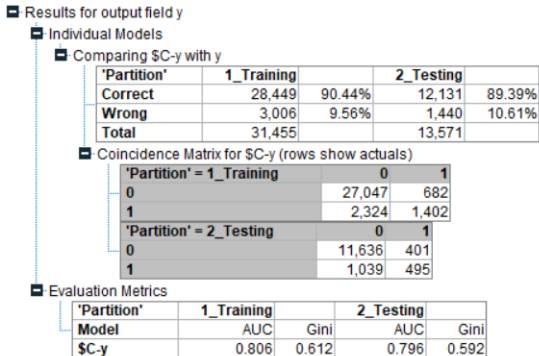
C5.0

- Drvo odlučivanja
- Tendencija je preciznost
- Dubina 13



Slika 5: Algoritam C5.0 - Važnost prediktora

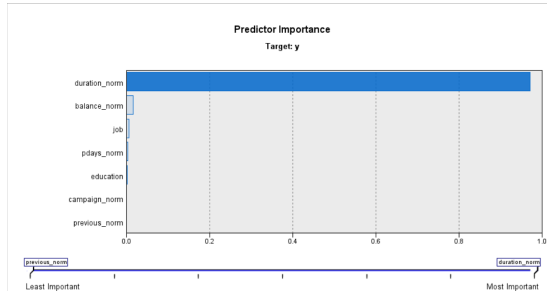
C5.0



Slika 6: Algoritam C5.0 - Matrica konfuzije

CART











- Minimalna dubina stabla: 4
- Ginijev indeks
- Bolja stabilnost - Enhance model stability
- Potrkivanje stabla radi izbegavanja prilagođenosti



Slika 7: Algoritam CART - Važnost prediktora

CART

Component Model Details

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

Slika 8: Algoritam CART - Modeli

CART

Results for output field y

Individual Models

Comparing \$R-y\$ with y

'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	

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

'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

Performance Evaluation

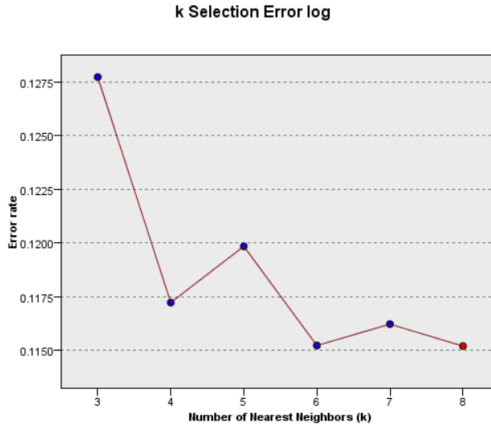
'Partition' = 1_Training	
0	0.021
1	1.595
'Partition' = 2_Testing	
0	0.02
1	1.637

Slika 9: Algoritam CART - Matrica konfuzije

KNN

- Cilj: Preciznost modela
- k između 3 i 8
- Menhetn rastojanje
- Prediktori jednako važni
- Za $k=6$ i $k=8$ najbolja rešenja

KNN



Slika 10: Algoritam KNN - Nivo greške u zavisnosti od k

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 11: Algoritam KNN - Matrica konfuzije

QUEST

- Binarno stablo
- Uvek dubina 2
- *duration* najbitniji
- Ostali prediktori jednako važni, ali zanemarljivi
- $AUC = 0.637$, blizu 0.5

QUEST

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\$

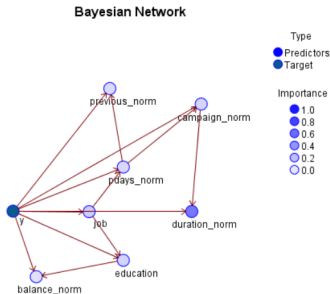
Evaluation Metrics

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

Slika 12: Algoritam QUEST - Matrica konfuzije

Bajesove mreže

- Bajesova statistika
- Posmatra *duration*, zatim *job*
- Uslovne verovatnoće svake instance



Slika 13: Algoritam Bajesove mreže

Bajesove mreže

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 14: Algoritam Bajesove mreže - Matrica konfuzije

Naivan Bajesov algoritam

- trening skup 0,75, test skup 0,25
- 1-yes, 0-no
- `sklearn.naive__bayes`
- Gausova formula

Naivan Bajesov algoritam

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

Izveštaj 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 15: Naivan Bajesov algoritam - Matrica konfuzije

Zaključak

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

Slika 16: Preciznost algoritama

Hvala na pažnji!